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,

38 as well as compare the profiles between multiple pathways.

39 When multiple genes are being examined together, the line plots are less effective for

40 visualizing DE or expression profiles: The lines often cross each other, making it difficult

41 to identify the grouping and understand the behavior of individual genes. One common

42 alternative visualization method is the scatterplot, which shows expression level under two

43 treatment conditions at a time. Scatterplots are effective in showing clustering patterns

44 and outliers, greatly assisting with data exploration (Elmqvist et al., 2008). For high

45 dimensional data, one has the option of using the scatterplot matrix, in which each panel

46 is the scatterplot for the corresponding pairs of features. However, manual scanning

47 of all possible pairwise scatterplots can be arduous or even fruitless at times, because

48 the number of possible visualization angles increases quadratically with respect to data

49 dimensionality (p choose 2 possible angles).

50 In this paper, we propose to automatically search for the best low dimensional visual-

51 ization angles (2-, 3-, or 4-dimensional) based on a context-sensitive, numeric measure of

52 importance, thereby reducing the amount of effort invested in scatterplot scanning. In our

53 study, we hope to explore the patterns and differences in gene expression profile between

54 two phytohormone signaling pathways, and therefore, we would like the top ranked scat-

55 terplots to contain as much information as possible on pathway classification. We thus

56 define such an importance measure for the dimension subsets such that the scatterplots

57 will show the largest separation between different pre-defined groups in the data set.

58 For this study, we will look for feature subsets upon which the pathways ethylene (ET)

59 and jasmonate (JA) are the most separated, and quantify the between-group separation by

60 calculating the repeated cross-validation (RCV) error of misclassification using MclustDA

61 (Fraley and Raftery, 2002), a model-based classification method. In Figure 1, we show

62 one of the top ranked 2-subset feature combinations that give the greatest separation

63 between two pathways, as well as subset giving the smallest separation. Comparing the

64 two scatterplots, we can observe that the two pathway groups in Figure 1(a) are more

65 visually distinguishable than those in Figure 1.
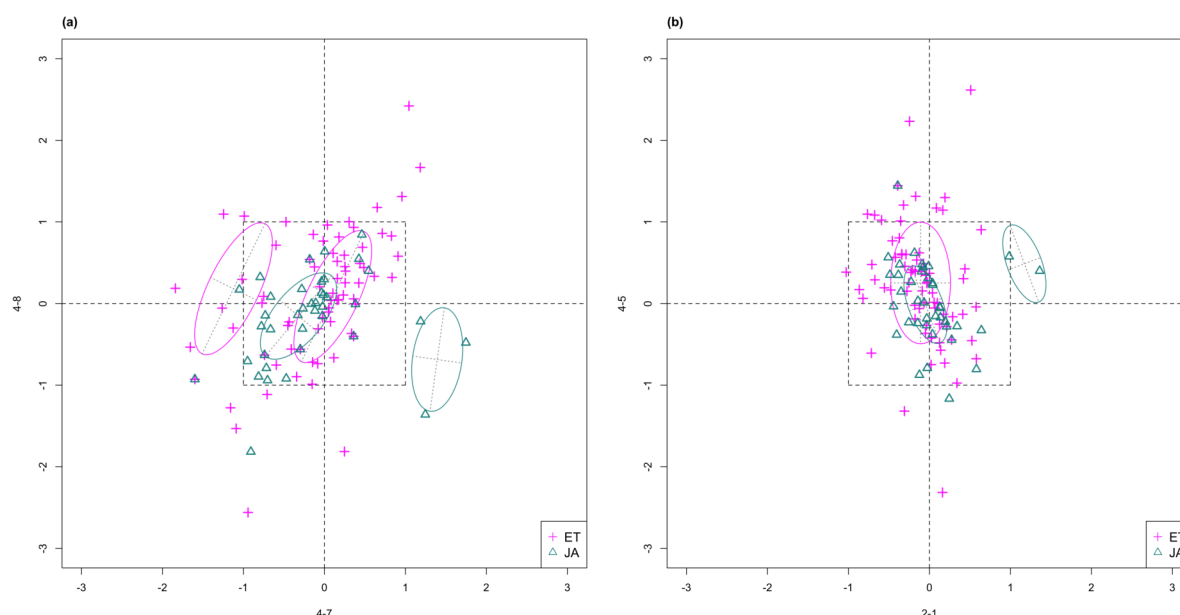
Figure 1: Scatterplots of 2-dim feature subsets reflecting maximum (A) and minimum (B) group separations. Dashed-line square marks ±1 range from the origin. Different classes distinguished with color. Ellipses correspond to component mean and covariance fitted by MclustDA. Treatment i-j represents the 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 a portion of the data collected and processed by Bin Zhuo (Zhuo et al., 2016). The original data 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

77 conditions (including genetic variants) varying between experiments. Tissue types in the

78 original data include leaf, seed, and multi-tissue. The number of treatments/factor levels

79 also vary among the experiments.

80 For this article, we will focus on experiments conducted on the leaf tissue of *Arabidop-*

81 *sis*, which include a total of 5 datasets. The GEO (Gene Expression Omnibus) accession

82 numbers (which can be used to directly search for the experiment/dataset information)

83 are available as part of the meta-data, and the assembly pipeline that produced the read

84 count matrix is described in detail in Zhuo et al. (2016). We have included some basic

85 information on the experiments in Table 1.

ID	GEO accession #	Title	Platform	Sample Size
1	GSE36626	Dynamic Deposition of the Histone H3.3 Variant Accompanies Developmental Remodeling of Arabidopsis Transcriptome (mRNA-Seq)	GPL11221 *Illumina Genome Analyzer IIx*	4
2	GSE39463	Time-course RNA-seq analysis of the barley MLA1 immune receptor-mediated response to barley powdery mildew fungus Bgh in *Arabidopsis thaliana*	GPL13222 *Illumina HiSeq 2000*	48
3	GSE48235	Four distinct types of dehydration stress memory genes in *Arabidopsis thaliana*	GPL9302 *Illumina Genome Analyzer II*	6
4	GSE51304	Non-CG methylation patterns shape the epigenetic landscape in *Arabidopsis*	GPL13222 *Illumina HiSeq 2000*	18
5	GSE54677	Transcriptional gene silencing by Arabidopsis Microrchidia homologues involves the formation of heteromers	GPL13222 *Illumina HiSeq 2000*	20

Table 1: Experiment information

2.2   Estimating log fold changes

86 Let c_{ij} be the raw count of RNA-Seq reads mapped to gene i in biological sample j. For

87 each gene i, to assess gene expression, we often consider the relative read frequencies $\frac{c_{ij}}{N_j}$

88 where N_j is the total number of reads mapped to sample j. N_j can vary greatly across

89 samples. A more subtle issue is that N_j can be significantly influenced by a few genes with

90 extremely high read counts. To address this issue, Anders and Huber (2010) proposed to

91 normalize N_j's by multiplying them with so-called normalization factors R_j's. Using their

92 method, R_j's are estimated such that the median fold change (over all genes i) will be

93 1 between each column of normalized relative frequencies, $\frac{c_{ij}}{N_j R_j}$, and a pseudo column,

94 $\frac{c_{i0}}{N_0 R_0}$ (for each gene i, c_{i0} is the geometric mean of c_{ij} over all samples $j = 1, ..., n$, N_0

95 is the column total of c_{i0}, and see Zhuo et al. (2016) for a formula for R_0 and other

96 details). In Zhuo et al. (2016), we identified a set of 104 genes that were shown to be

97 relatively stably expressed across all biological samples in a collection of *Arabidopsis* RNA-

98 Seq experiments (including the 5 experiments we use in this study). In this paper, we

99 used these 104 genes as a reference gene set to compute the normalization factors. With

100 the normalization factors R_j estimated this way, the median forld change in normalized

101 relative frequencies over these 104 genes will be 1 between any sample j and the pseudo

102 sample.

103   With these estimated normalization factors R_j, we fit a log-linear negative binomial

104 regression model to each row i (one gene) of the read count matrix (separately for each

105 experiment) to assess differential expression between treatment groups (we suppressed the

106 subscript i in the model equation below):

$$c_j \sim \text{Negative Binomial}(\mu_j, \phi),$$

$$\log(\mu_j) = \log(N_j R_j) + x_{j0}\beta_0 + x_{j1}\beta_1 + ... + x_{jp}\beta_p,$$

107 for $j = 1, ..., n$, where ϕ is the overdispersion parameter capturing the commonly oberved

108 extra-Poisson variation in the read count data, and (x_{jk}) is the model matrix. We chose

109 the model matrix such that the fitted value of β_0 will correspond to the log mean relative

110 count of the control group, and the fitted regression coefficients $\beta_1, ..., \beta_p$ will correspond

111 to log fold changes between a treatment group and the control group. We used our own

112 NBPSeq package to fit such a log-linear regression model (Di, 2015). These estimated log

113 fold changes (converted to base 2) from the 5 experiments described in Section 2.1 are

114 the feature sets used in this paper.

2.3   Finding pathway genes

115 For this study, we focus our attention on the signaling pathways of two phytohormones:

116 ethylene (ET) and jasmonic acid (JA). As a plant hormone, ethylene is commercially

117 important due to its regulation on fruit ripening (Lin et al., 2009). JA acts as a key

118 cellular signal involved in the activation of immune responses to most insect herbivores

119 and necrotrophic microorganisms (Ballaré, 2010).

120   For each pathway, we first use AmiGO 2 (`http://amigo.geneontology.org/amigo/`

121 `landing`) to search for the list of genes involved, and then identify the subset of genes

122 in our data set that belong to the pathway through cross-reference. Genes with a fold

123 change of < 2 under all treatment conditions are filtered out. The name, GO accession

124 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

125 In Figure 2, we display the expression profile of genes that belong to each pathway

126 group. Under certain individual treatment-control contrasts (e.g. 2-3, 4-3, 5-1), there

127 exist observable similarities between the distribution of expression levels, while it is more

128 difficult to tell under other treatments.
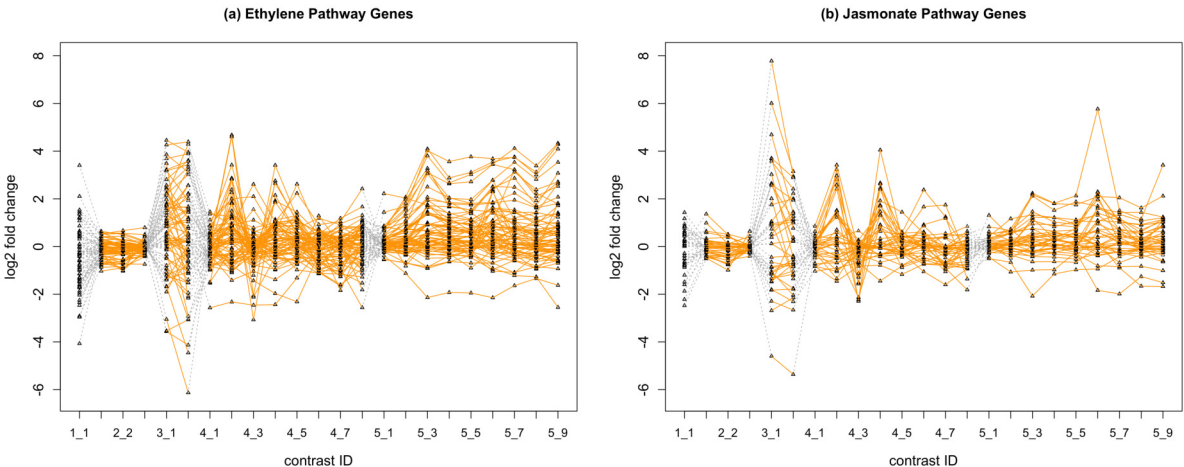
Figure 2: Gene expression profile plot for pathways ET (A) and JA (B). Treatments from the same experiment are joined by orange lines. Different experiments are joined by grey dashed lines. Feature i-j represents the jth treatment-control contrast in experiment i.

3. METHOD

3.1 Mixture discriminant analysis via MclustDA

129 In this section, we will start by introducing a classification method named MclustDA, and

130 then define a measure for group separation using cross-validation results with MclustDA.

131 Finally, we lay out our strategy for reducing data dimensionality with the ultimate goal

132 of simplifying navigation of scatterplots.

133

134 **MclustDA model**   In discriminant analysis (DA), known classifications of some ob-

135 servations are used to classify others. The number of classes, G, is assumed to be known.

136 For probabilistic DA methods, it is assumed that observations in class k follow a class

137 specific probability distribution $f_k(\cdot)$. Let τ_k represent the proportion of observations in

138 class k. According to Bayes's theorem, it follows that

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

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

140 Commonly used DA methods, including Fisher's linear discriminant analysis (LDA)

141 and quadratic discriminant analysis (QDA), assume a multivariate normal density for

142 each class:

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

143 The method is called LDA if the covariance matrices for all classes coincide ($\Sigma_k = \Sigma$

144 for $k = 1, ..., G$), and is called QDA if the class covariances are allowed to vary.

145 MclustDA (Fraley and Raftery, 2002), an extension and generalization to LDA and

146 QDA, models each class density as a mixture of multivariate normals. The density for

147 class j is as follows:

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

148 where G_j is the number of components for class j, $\{\tau_{jk}\}$ are mixing proportions for com-

149 ponents in class j, and θ_k is the vector of parameters for the normal mixture. Component

150 covariances Σ_{jk} are allowed to vary both within and between classes.

151 Parameters within each class are separately estimated by maximum likelihood via the

152 EM algorithm (Dempster et al., 1977), which is equivalent to fitting a Mclust (Fraley and

153 Raftery, 2002) model for each class. And just like Mclust, MclustDA performs model

154 selection within each class for the number of mixture components as well as covariance

155 matrix parameterizations with Bayesian information criterion (Schwarz, 1978).

156

157 **Comparison with LDA**   In our study, MclustDA is chosen over LDA/QDA as the

158 classifier due to its greater flexibility in describing the data. In RNA-Seq analysis, we

159 typically assume that the majority of genes are not differentially expressed, and therefore

160 we expect to see a cluster of points around the origin. Since MclustDA proposes to fit

161 more than one normal component to each class, it readily captures the cluster of non-DE

162 genes as well as any abnormalities that might be of interest.

163   In Figure 3, we fitted a MclustDA model and a LDA model on dimensions [3-1, 4-

164 2] of our data, separately. In MclustDA fit, each class is described with a mixture of

165 two bivariate normal components, with the ellipses representing fitted covariance matrix

166 estimates. For details in how the ellipses are constructed, see Appendix A.

167   Class JA is fitted with a component centered near the origin, representing genes

168 with low expression levels under both treatments, as well as a component centered at

169 $(2.276, 1.663)$ that encompasses relatively active genes. Class ET is represented by a

170 single normal component centered at $(0.537, 0.406)$.

171   In comparison, due to model assumptions, LDA fitted a bivariate normal density to

172 each class with covariances being equal, and in this case, the estimated centers almost

173 coincide with each other. The fitted normal densities are only able to capture the general

174 shape and orientation of each class, while MclustDA provides us with a more detailed

175 anatomy of geometric and distributional properties in each class.
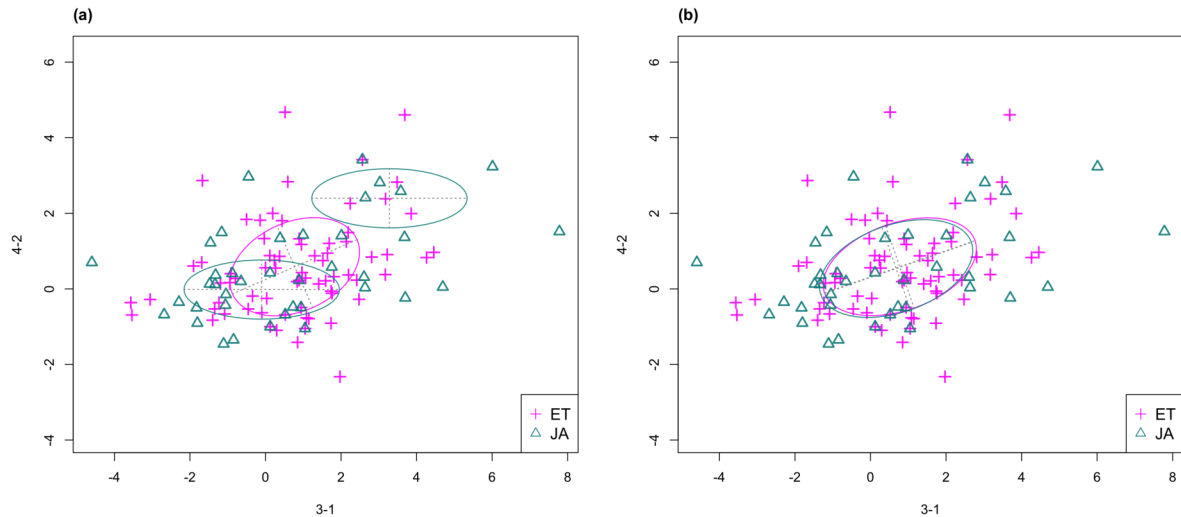
Figure 3: Comparison of MclustDA (A) and LDA (B) fit of the same data. Fitted components and points from different classes are distinguished with color. Ellipses correspond to component covariances.

3.2 Quantification of group separation

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

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

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

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

186 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.

$$\text{TPM} = p_1 P(2|1) + p_2 P(1|2). \tag{1}$$

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:

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.

227   Let $C \times K$-CV denote a K-fold CV with C repetitions. There has been much discus-

228 sion on the optimal choice of C and K (Kohavi, 1995; Kim, 2009; Powers and Atyabi,

229 2012). Increasing C tends to decrease the variance of the RCV estimator, but at the same

230 time increases the computational time. The choice of K takes into account the tradeoff

231 between bias and variance of the CV estimator (of the expected error rate). For small

232 K, less data is used to train the classifier and therefore the error estimate tends to be

233 biased. For large K, the estimator becomes less biased due to more data being used in

234 training, but its variance is inflated due to higher correlation between different training

235 folds. Kohavi (1995) recommends using a stratified 10-fold CV with multiple runs, and

236 we chose $C = 10$ considering the amount of computation required as well as the specs of

237 our hardware.

238

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

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

240 where $\hat{\epsilon}_{\mathrm{rcv}}$ denotes the repeated stratified CV estimator of the total misclassification prob-

241 ability using MclustDA as the classifier.

242   Intuitively, for a chosen feature subset, a small CV error indicates that the data can be

243 more easily classified when projected onto these dimensions, which, in our expectation,

244 can be reflected in the graphical representation of the data by showing that different

245 classes can be more easily distinguished through simple visualization.

3.3   Feature subset selection via GSI ranking

246 In this section, we describe the data in each pathway with a low dimensional representation

247 for easier interpretation by selecting a parsimonious subset of features (treatment-control

248 contrasts) that contain as much information on pathway classification/separation as pos-

249 sible. In other words, we hope to find the dimensions to project the data onto such that

250 the separation between two pathways is as large as possible. We use GSI, as defined in (3),

251 to measure the separation between two pathway groups.

252   In order to find the optimal feature subset in terms of group separation, we designed

253 and implemented the following algorithm:

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

255 **Step 2**: List all M-subsets of features exhaustively. Call this collection of subsets

256 $\mathcal{F}_M$.

257 **Step 3**: For each member of $\mathcal{F}_M$, subset the data accordingly. Calculate and

258 record a 10×10 stratified CV error rate (and equivalently, GSI) with MclustDA

259 classifier on each subsetted data. For each fold of CV, use misclassification error

260 rate as measure of fit.

261 – CV model fitting: First fit a MclustDA model to the entire subsetted data,

262 setting maximum number of components as $G_j \equiv G = 2$. Then use the same

263 fitted model (number of components, covariance parameterization) for every

264 fold of CV.

265 **Step 4**: Rank the feature subsets in $\mathcal{F}_M$ according to their GSI values. Feature

266 subsets with higher GSI values are ranked higher.

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

268 Concerning the choice of maximum number of components in Step 3, we intially at-

269 tempted to use the default value $G = 5$ in R function `MclustDA`, and discovered that in

270 some situations the number of model parameters became too large for the algorithm to

271 produce a meaningful point estimate. The same problem occurred for $G = 4$ and $G = 3$,

272 especially when we looked at 3- and 4-dimensional data. Therefore, we settled on using

273 $G = 2$ for our implementation.

274 For the purpose of finding "good" angles for data visualization, we will examine the

275 scatterplots and scatterplot matrices generated by top-ranked feature subsets. The results

276 will be discussed in Section 4.

277

278 **Random number seed** To ensure reproducibility of our results, for each of 2-, 3-

279 and 4-subset selection process, we followed the following protocol to set random number

280 seeds:

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

282 of 10 seeds.

283 **Step 2**: For each feature subset, run 10-fold stratified CV for 50 times, each time

284 using a different seed from the list.

285 **Step 3**: Average results within each of 5 batches of 10 random seeds to obtain

286 10×10 stratified CV result. For instance, average of seeds $1 \sim 10$ results serves as

287 first run of 10×10 RCV; average of seeds $11 \sim 20$ serves as second run, etc.

3.4   GSI ranking with LDA as classifier

288 As discussed in Section 3.1, LDA fits only one multivariate normal component to ob-

289 servations in each class. In gene expression data such as ours, a majority of genes are

290 expected to be expressed at a low level, meaning we are likely to observe a cluster of data

291 points around the origin, regardless of the class they belong to. Meanwhile, any aberrant

292 patterns demonstrated by individual data points are often not captured by fitting LDA

293 model.

294   To see whether our method can still discover interesting visualization angles with LDA

295 as classifier, we modified and implemented the GSI calculation algorithm accordingly, and

296 applied it to our data. Results are presented in Section 4.2.

3.5   Dimension reduction via PCA

297 Principal component analysis (PCA) maps the data onto a lower dimensional space in such

298 a way that the variance of the data in the low-dimensional representation is maximized.

299 As a dimension reduction technique, usually only the first few principal components (PCs)

300 are used. Despite its popularity in the field of data visualization, the formulation of PCA

301 does not involve any class information in the data, which implies that the projected

302 directions corresponding to the largest variance may not contain the best separability

303 information.

304   To verify this observation, using the expression data from all 5 experiments, we calcu-

305 lated its principal components, and treat them as the new (projected) features. Then for

306 the first 2, 3 and 4 PCs, respectively, we calculated the group separation index for each

307 case using 10×10-CV with MclustDA and compare the results with ours.

4. RESULTS

4.1 Repeated cross-validation with MclustDA

308 With the secondary purpose of testing the stability of repeated CV, we executed multiple

309 runs for each of the 2-, 3-, and 4-subset feature selection procedures. The top ranked

310 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:

313 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]

314 are among top ranked feature combinations in all 5 runs.

315 For 3-subset (Table 5), feature combinations [3-2, 4-7, 5-2], [3-2, 5-2, 5-6] and [1-1,

316 3-1, 5-2] are ranked top for all 5 runs.

317 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

318 ranked feature combinations for all runs.

4.1.2 Top Ranked Scatterplot: Same Experiment

319 In Figure 4, we show the scatterplot of the data projected onto dimensions [4-7, 4-8], one

320 of the top ranked 2-subset feature combinations. These two features originate from the

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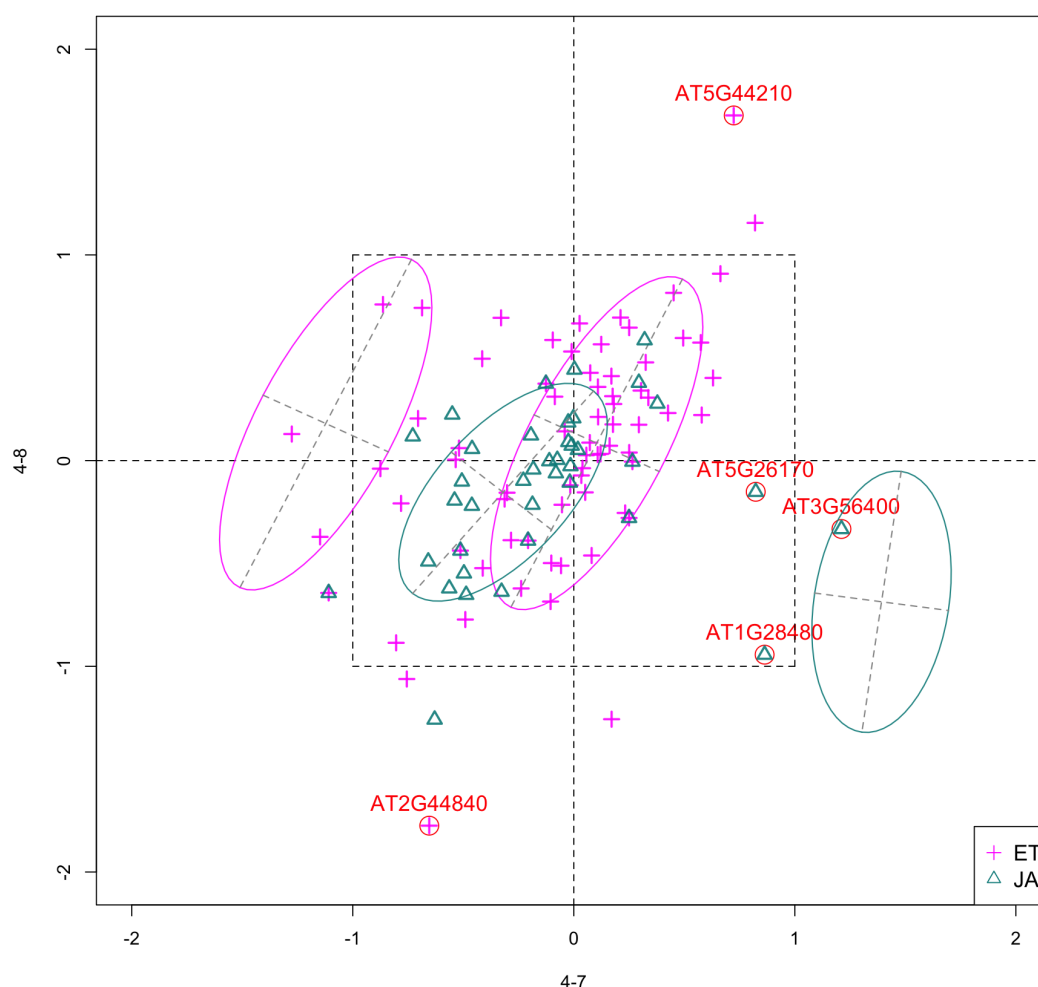
Figure 4: Scatterplot of data projected on dimensions 4-7 and 4-8. Pathways are distinguished with color. Ellipses represent estimated covariances fitted by MclustDA. Potential outliers highlighted and labeled with their names. Dashed-line square is $\pm \log(2)$ range from the origin.

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 chromomethy-

330 lase (CMT) and domains rearranged MTase (DRM) proteins, in particular CMT3 and

331 DRM2. Expression of DRM1 is scarcely detected, while the function of CMT2 has not

332 been studied as well as that of CMT3.

333 Histone methylation is a process by which methyl groups are transferred to amino acids

334 of histone proteins. Histone methylation can either increase or decrease gene transcription,

335 depending on which amino acids are methylated and the degree of methylation. The

336 methylation process is most commonly observed on lysine residues (K) of histone tails H3

337 and H4, among which H3K9 (lysine residue at 9th position on H3) serves as a common site

338 for gene inactivation. Lysine methylation requires a specific MTase, usually containing an

339 evolutionarily conserved SET domain. In *Arabidopsis*, Su(var)3-9 homologue 4 (SUVH

340 4), SUVH 5 and SUVH 6 are the major H3K9 MTases.

341 Feature 4-7 corresponds to the *drm1 drm2 cmt2 cmt3* quadruple gene knockout mu-

342 tant, created by crossing *cmt2* to *cmt3* and *drm1 drm2* double mutants. It was found

343 that non-CG methylation was eliminated in such mutants, indicating that DRM1, DRM2,

344 CMT2 and CMT3 proteins are collectively responsible for all non-CG methylation in *Ara-*

345 *bidopsis*. Feature 4-8 corresponds to the *suvh4 suvh5 suvh6* triple mutant. The control

346 group of this experiment corresponds to wildtype *Arabidopsis*. Table 7 summarizes the

347 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

348 **Outliers** Potential outliers from JA pathway, as highlighted and labeled in the scat-

349 terplot, fall into the fourth quadrant, which implicates that these genes are up-regulated

350 under 4-7 (DNA methylation) but down-regulated under 4-8 (histone methylation). In-

351 formation on these genes is collected from TAIR (Berardini et al., 2015) and displayed in

352 Table 8. One interesting discovery we made was that one of the outliers, **AT3G56400**,

353 functions as a repressor of JA-regulated genes. In other words, its gene product inhibits

354 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

370 In Figure 5, we show the scatterplot of another top ranked feature combination, [3-1, 5-2],

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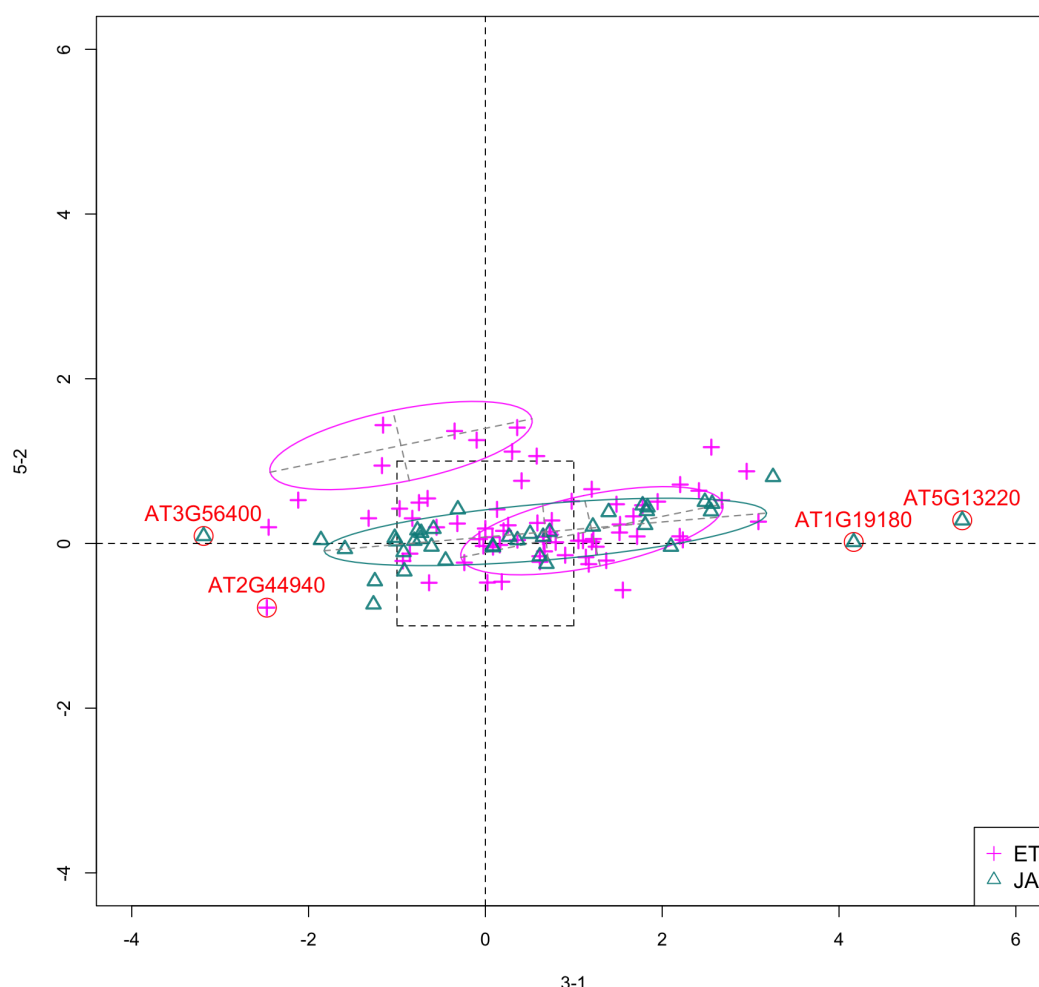
Figure 5: Scatterplot of data projected on dimensions 3-1 and 5-2. Pathways are distinguished with color. Ellipses represent estimated covariances fitted by MclustDA. Potential outliers highlighted and labeled with their names. Dashed-line square is $\pm \log(2)$ range from the origin.

372 **Experiment 3** The focus of this study is the response of *Arabidopsis* to multiple

373 consecutive dehydration stresses (Ding et al., 2013). Based on the observation that pre-

374 exposure to abiotic stresses (including dehydration) may alter plants subsequent responses

375 by improving resistance to future exposures, the researchers hypothesized the existence of

376 "memory genes": genes that provide altered response to subsequent stresses (Ding et al.,

377 2012).

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

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

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

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

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

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

392

393 **Experiment 5**   In this study, the researchers examine the functional relationship be-
394 tween members of the *Arabidopsis* microchidia (AtMORC) ATPase family (Moissiard
395 et al., 2014), which have been shown to be involved in transposon repression and gene
396 silencing. Three of seven MORC homologs were examined: AtMORC1, AtMORC2 and
397 AtMORC6. RNA-Seq experiment using single and double mutants indicates that At-
398 MORC1 and AtMORC2 act redundantly in gene silencing. Wildtype *Arabidopsis* was
399 used as control group. Treatment groups include both single and double mutant lines:
400 *atmorc2-1*, *atmorc2-4*, *atmorc1-2*, *atmorc1-5*, and *atmorc1-2 atmorc2-1*, in which two in-
401 dividual alleles were used for *atmorc1* and *atmorc2*. In our data, feature 5-2 corresponds
402 to the single mutant line *atmorc2-1*. Table 9 includes summary information on this ex-
403 periment.

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

404 **Outliers**   In Figure 5, we highlighted a few observations considered as outlying, and

405 as before, looked up their information using TAIR. A brief description for each outlier is

406 included in Table 10. Gene **AT3G56400** is again identified as an outlier, mainly because

407 of its highly negative expression level under treatment 3-1, while the near-zero expression

408 level under 5-2 indicates its inactivity under this treatment. Gene **AT5G13220** has the

409 highest expression level under 3-1 among all JA genes, and at the same time not as active

410 under 5-2. This gene is interesting because it functions as a repressor of JA signaling,

411 and its high expression level could be an implication for repression of JA singaling for

412 *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

413 **Pattern differences**   From the scatterplot, the first thing we can observe is that quite

414 a few genes from both pathways are up- or down-regulated under treatment 3-1, while

415 genes are expressed at an overall low level under 5-2. Nevertheless, a few genes from ET

416 group show overexpression pattern under 5-2. JA pathway genes populate quadrants I,

417 II and III, while ET pathway genes are mainly located in quadrants I, II and IV. Overall,

418 under 5-2, ET genes tends to be more active than JA genes.

4.2   Using Fisher LDA as classifier

419 With Fisher's LDA as the classifier, we found a distinct assembly of top ranked feature

420 pairs than when MclustDA was used. Table 11 show which 2-subsets produced higher

421 GSI scores than most others.

Table 11: 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	[3-1, 5-9]	0.700
2	[3-1, 5-5]	0.688
3	[4-4, 4-5]	0.688
4	[2-3, 3-1]	0.679
5	[3-1, 5-3]	0.675

(b) Run 2

Rank	Subset	GSI
1	[3-1, 5-9]	0.695
2	[3-1, 5-5]	0.687
3	[4-4, 4-5]	0.687
4	[3-1, 5-3]	0.682
5	[3-1, 4-5]	0.678

(c) Run 3

Rank	Subset	GSI
1	[3-1, 5-9]	0.704
2	[3-1, 5-5]	0.693
3	[4-4, 4-5]	0.688
4	[3-1, 5-3]	0.683
5	[3-1, 5-1]	0.675

(d) Run 4

Rank	Subset	GSI
1	[3-1, 5-9]	0.700
2	[4-4, 4-5]	0.689
3	[3-1, 4-5]	0.680
4	[3-1, 5-3]	0.680
5	[3-1, 5-5]	0.680

(e) Run 5

Rank	Subset	GSI
1	[3-1, 5-9]	0.696
2	[3-1, 5-5]	0.693
3	[4-4, 4-5]	0.685
4	[2-3, 3-1]	0.683
5	[3-1, 4-5]	0.679

In Figures 6 and 7, we show the scatterplot of data projected on to dimensions [3-1, 5-5] and [3-1, 5-9], two of the top ranked feature pairs. Visually, the two pathways groups are not as evidently separated as seen in our previous examples (Figure 4 and 5).
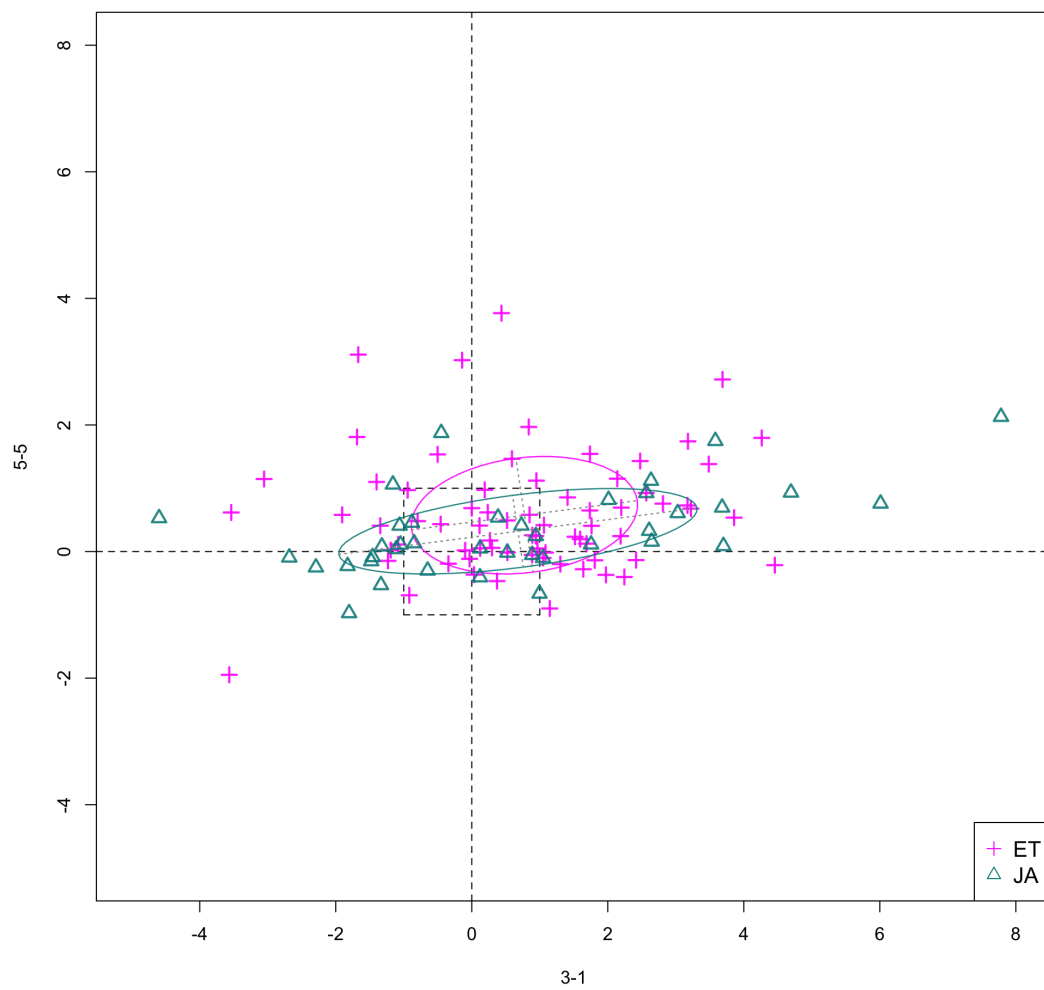
Figure 6: Scatterplot of data projected on dimensions 3-1 and 5-5.
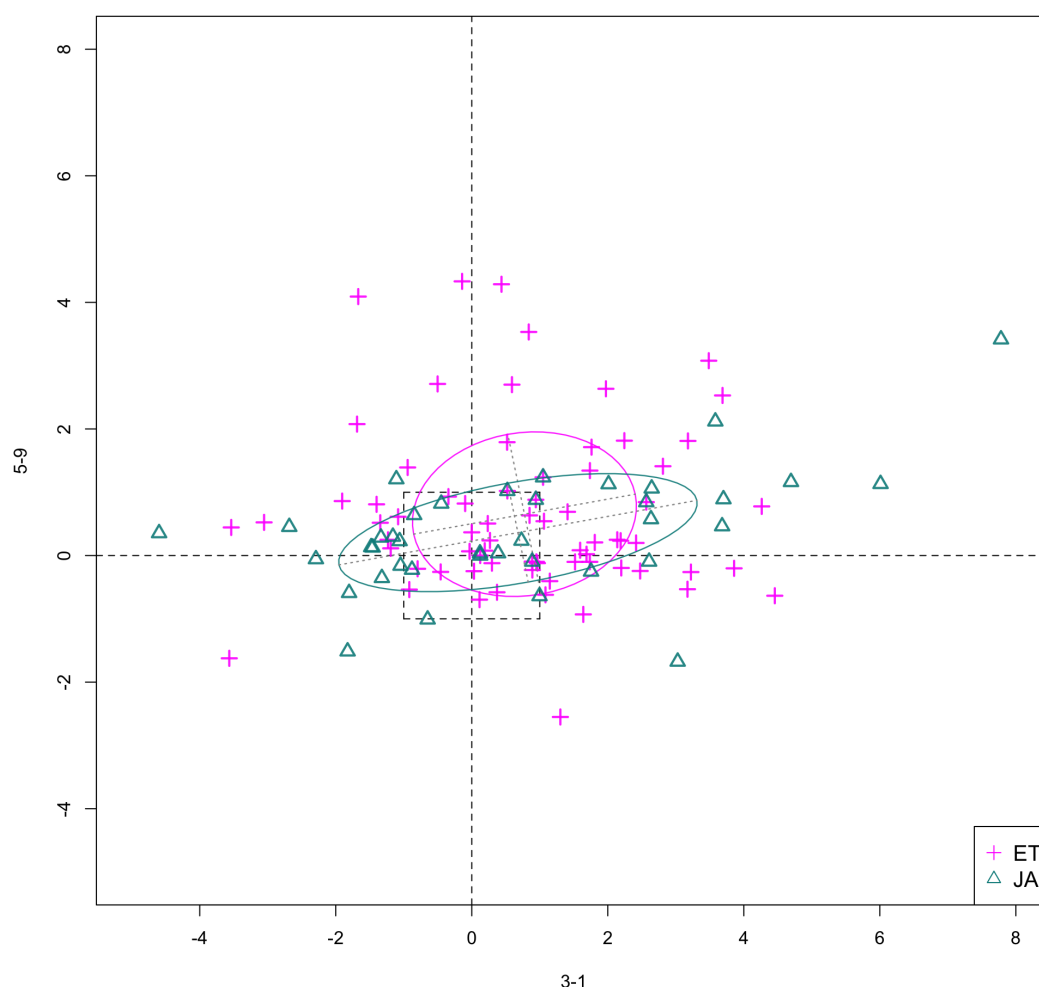
Figure 7: Scatterplot of data projected on dimensions 3-1 and 5-9.

4.3   GSI for PC transformed data

In Table 12, 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 3.5.

Table 12: 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 of work   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 13), 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. In Section 4.1, we singled out two of the top ranked scatterplots for discussion. Interested readers are directed to the appendix for additional scatterplots and scatterplot matrices for top ranked 3- and 4-subsets (Figures 8~15).

Table 13: 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

Moreover, as pointed out by one of the reviewers, as the sample size increases, one can hope that the bias of the error (hence GSI) estimation will reduce and the ranking of feature subsets from each CV run will become more stable. One can then save computational time by eliminating the repetition, or even the use of cross-validation in the calculation of GSI. In the Appendix, we illustrated how the bias in error estimation will change and how the variability of GSI values from repeated CV will decrease as the sample size increases using simulated data.

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

475 by only looking at individual panels of scatterplots. To more effectively visualize higher

476 dimensional feature subsets, we can consider using interactive visualization tools, such as

477 GGobi (Swayne et al., 2003) and R Shiny (Chang et al., 2017). Both tools allow users

478 to identify the same point in all panels of a scatterplot matrix, significantly increasing its

479 visual expressiveness.

480

481 **Error rate definition**   In our definition of TPM in (1), we made the assumption that

482 the cost of misclassifying an object from either class is the same, i.e. $c(1|2) = c(2|1)$. We

483 can adjust the cost values if we are more concerned about correctly classifying a certain

484 class of observations.

485

486 **Evaluating reproducibility of experiments**   Currently, a typical differential expres-

487 sion analysis is conducted in a gene-wise manner, i.e. genes are treated as observations and

488 the treatment conditions as features. In our study, we took the same approach because

489 our goal was to differentiate expression pattern between two groups of genes. However,

490 with the increase in the availability of RNA-Seq data thanks to advances in information

491 technology, we can also study the comparability and reproducibility of RNA-Seq experi-

492 ments. In this sense, we will be exploring the relationship between treatment conditions

493 or experiments, with genes acting as features/variables. Evaluation of experiment repro-

494 ducibility is usually accomplished by performing the same experiment using the same

495 setting, which is, unfortunately, not a common practice in RNA-Seq studies. In light of

496 this, one of our long-term goal is the quantification of similarity between RNA-Seq ex-

497 periments, which not only accounts for differences in experimental designs and parameter

498 settings, but also utilize the information hidden in the expression of genes.

499

500 **Significance evaluation**   One reviewer commented on whether we can assign a mea-

501 sure of significance for the top ranked feature set found by our method. Our interpretation

502 of "significance" in this context is as follows: Under the assumption that the feature pair

503 contain no group information on the genes, what is the probability of obtaining a GSI

504 value as high as or higher than the value computed by our method (i.e. a p-value)? We

505 can use a permutation test for this purpose, by following these steps: (1) For a particu-

506 lar feature pair, randomly reassign group labels to genes, (2) Calculate GSI value using

507 the permuted data, (3) Repeat 1 and 2 for M (usually large) times, and establish the
508 distribution of GSI value, and (4) Calculate empirical p-value based on this distribution.
509 Due to time constraint, we found such a p-value for only one top ranked feature pair,
510 [4-7, 4-8], by calculating $M = 2000$ GSI values from permuted data, and obtained the
511 empirical p-value of 0.0055, indicating the significance of this pair of features.

512

513 **Handling effect & CV**   As shown in Section 3.2, cross-validation was used for the
514 calculation of misclassification error rate. A recent study (Qin et al., 2016) points out
515 that in the existence of handling/batch effect of objects to be classified, CV will often
516 underestimate the error rate, mainly due to the complete or partial confounding between
517 batches and subject classes. In our study, the biological samples were indeed handled
518 separately under different circumstances, which means the treatment-control constrasts
519 derived using these samples are also subject to an inherent group structure. However,
520 rather than classifying the biological samples or contrasts, this study focuses on classifying
521 genes to one of two pre-specified gene pathway groups. Since a common assembly pipeline
522 was used to process the data, the list of genes should be considered as uniformly handled,
523 and therefore free from the handling/batch effect discussed in the aforementioned article.
524 The batch effect will, on the other hand, affect the covariance structure of the space of
525 log fold changes that we explore, which will be reflected when mclust models are fitted to
526 the lower dimensional subspaces.

6.   ACKNOWLEDGMENTS

527 The authors gratefully acknowledge Jeff Chang, Sarah Emerson and Duo Jiang for their
528 valuable insight and comments, and Bin Zhuo for the collection of data and the discovery
529 of a set of stably expressed reference genes.

REFERENCES

530 Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count
531   data. *Genome Biology*, 11(R106), 2010.

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

534 Tanya Z. Berardini, Leonore Reiser, Donghui Li, Yarik Mezheritsky, Robert Muller, Emily
535   Strait, and Eva Huala. The Arabidopsis Information Resource: Making and mining the
536   "gold standard" annotated reference plant genome. *genesis*, 2015. doi: 10.1002/dvg.
537   22877.

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

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

544 Yanming Di. Single-gene negative binomial regression models for RNA-Seq data with
545   higher-order asymptotic inference. *Statistics and Its Interface*, 8:405–418, 2015.

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

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

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

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

557 Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated
558   hold-out and bootstrap. *Computational Statistics and Data Analysis*, 53(11):3735–3745,
559   2009.

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

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

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

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

571 Li-Xuan Qin, Huei-Chung Huang, and Colin B. Begg. Cautionary note on using cross-
572   validation for molecular classification. *Journal of Clinical Oncology*, 34(32):3931–3938,
573   2016.

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

575 Hume Stroud, Truman Do, Jiamu Du, Xuehua Zhong, Suhua Feng, Lianna Johnson,
576   Dinshao J. Patel, and Steven E. Jacobsen. Non-CG methylation patterns shape the
577   epigenetic landscape in *Arabidopsis*. *Nature Structural & Molecular Biology*, 21(1):
578   64–72, 2014.

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

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

A. CONSTRUCTION OF COVARIANCE ELLIPSES FOR NORMAL COMPONENTS

584 In this section, we introduce how the covariance ellipses are constructed by MclustDA

585 when a scatterplot or a scatterplot matrix is graphed.

586

587 For 2D data, suppose the mean and covariance estimates for component k of class j

588 are $\hat{\mu}_{jk}$ and $\hat{\Sigma}_{jk}$, respectively. Also suppose that $\hat{\Sigma}_{jk}$ has eigenvalues $\lambda_1 \geqslant \lambda_2$ and their

589 corresponding eigenvectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$. Then MclustDA computes the major and minor

590 axes of the ellipse centered at $\hat{\mu}_{jk}$ the following way:

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

591 and the resulting ellipse has coverage probability of approximately 0.393.

592

593 In the case of higher dimensional data, MclustDA constructs the scatterplot and

594 graphs the ellipses two dimensions at a time. Suppose $\hat{\mu}_{jk}$ and $\hat{\Sigma}_{jk}$ are defined the same

595 way as above, and consider data dimensions p and q for visualization via scatterplot.

596 Let $\Sigma^{(p,q)} = [\hat{\Sigma}_{jk}]_{(p,q)}$ be the covariance submatrix corresponding to the two dimensions,

597 and $\mu^{(p,q)} = [\hat{\mu}_{jk}]_{(p,q)}$ be the corresponding mean vector. Now, suppose $\Sigma^{(p,q)}$ has eigen-

598 value/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

599 ellipse plotted by MclustDA has major and minor axes as follows:

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

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

B. SCATTERPLOTS AND SCATTERPLOT MATRICES FOR SELECT TOP RANKED FEATURE COMBINATIONS
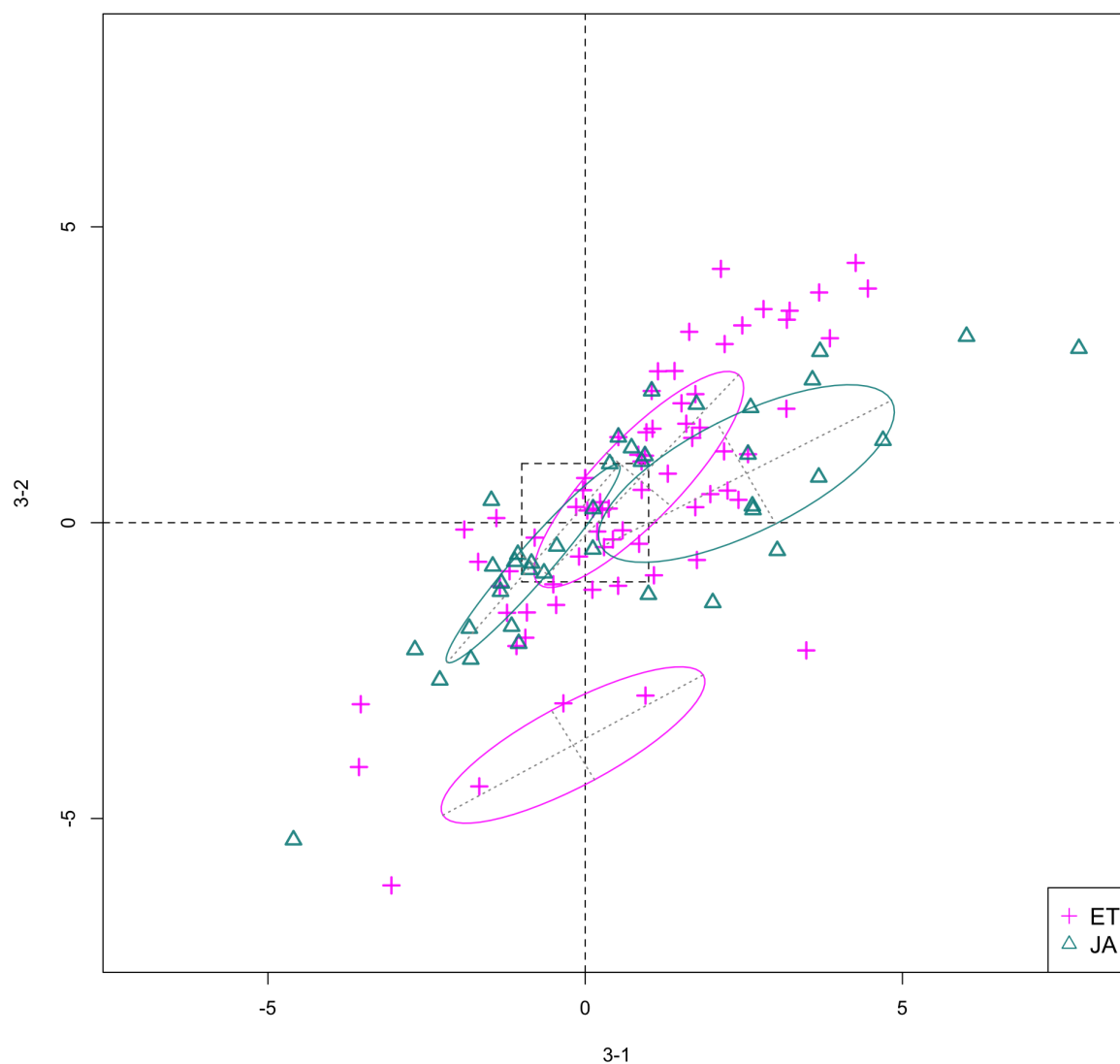
Figure 8: Scatterplot for 3-1 and 3-2
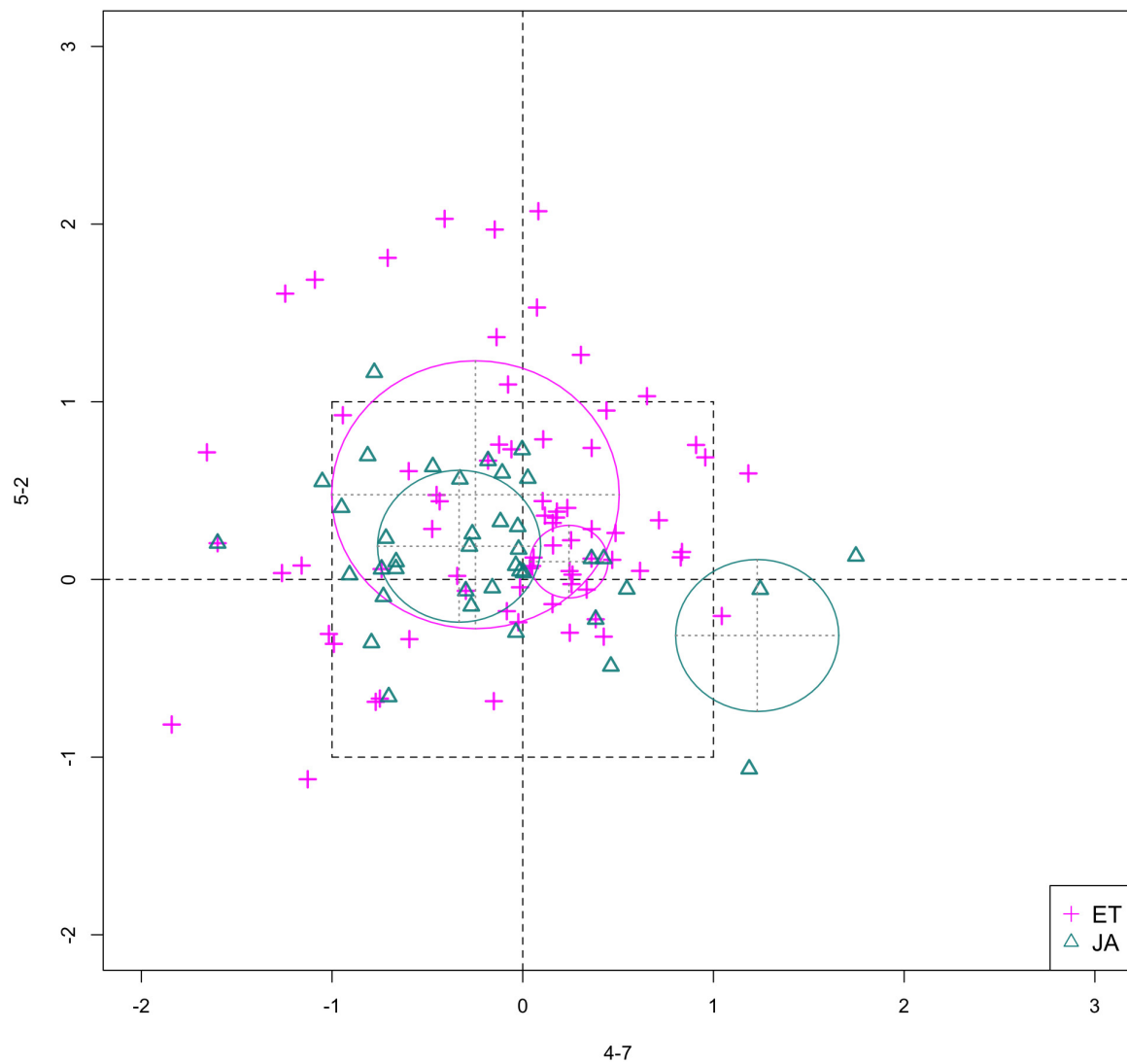
Figure 9: Scatterplot for 4-7 and 5-2
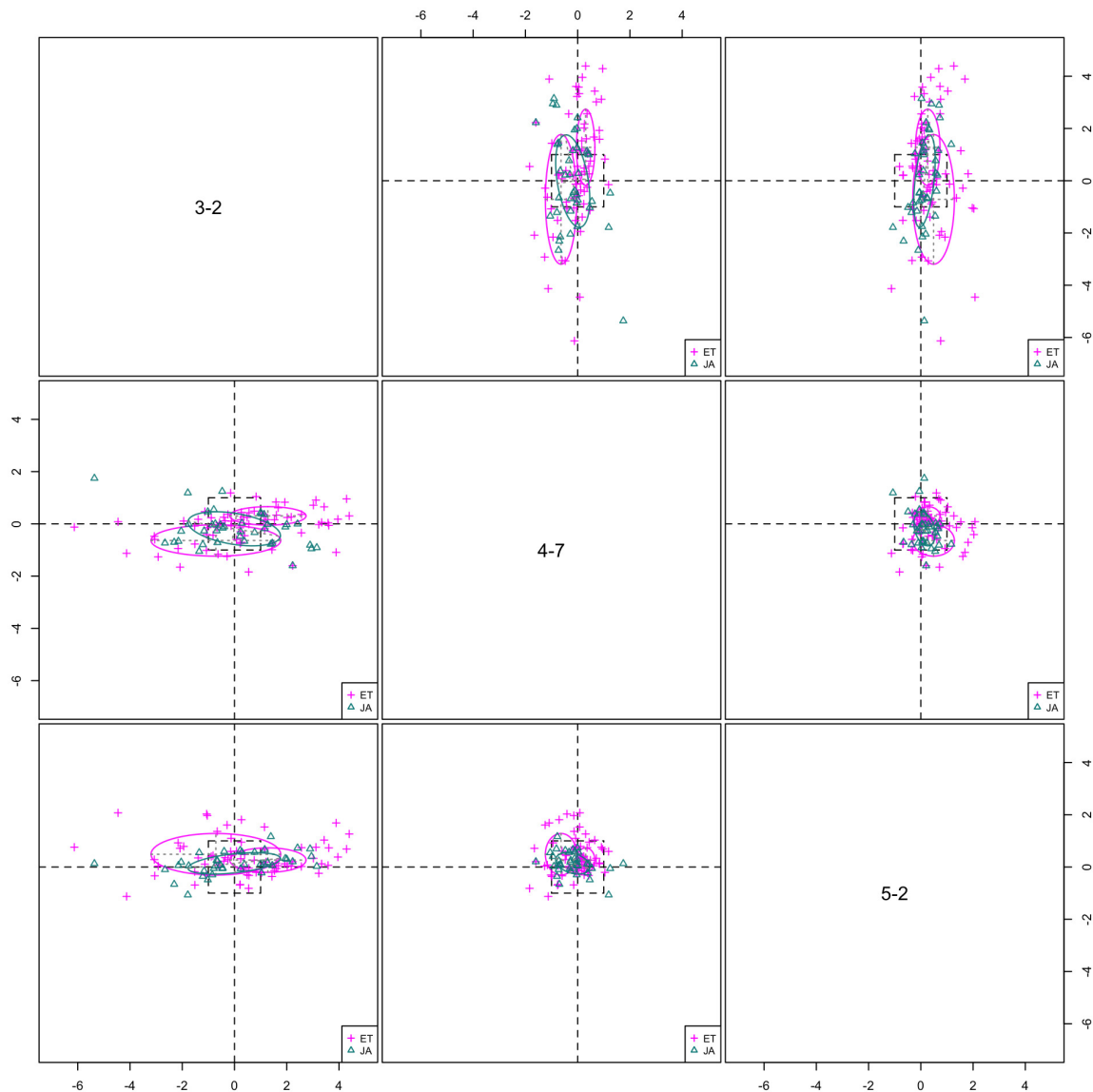
Figure 10: Scatterplot matrix for [3-2, 4-7, 5-2]
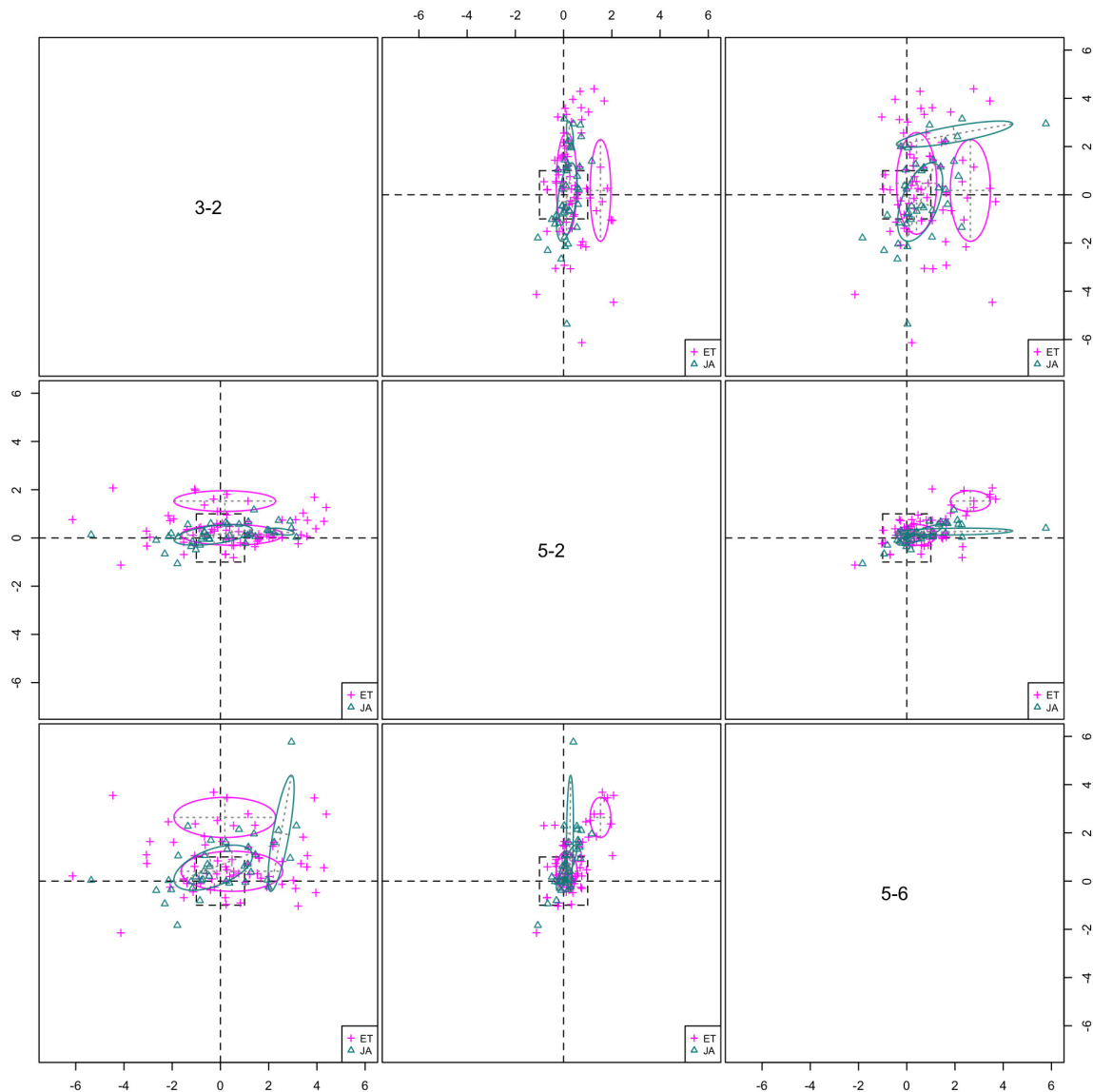
Figure 11: Scatterplot matrix for [3-2, 5-2, 5-6]
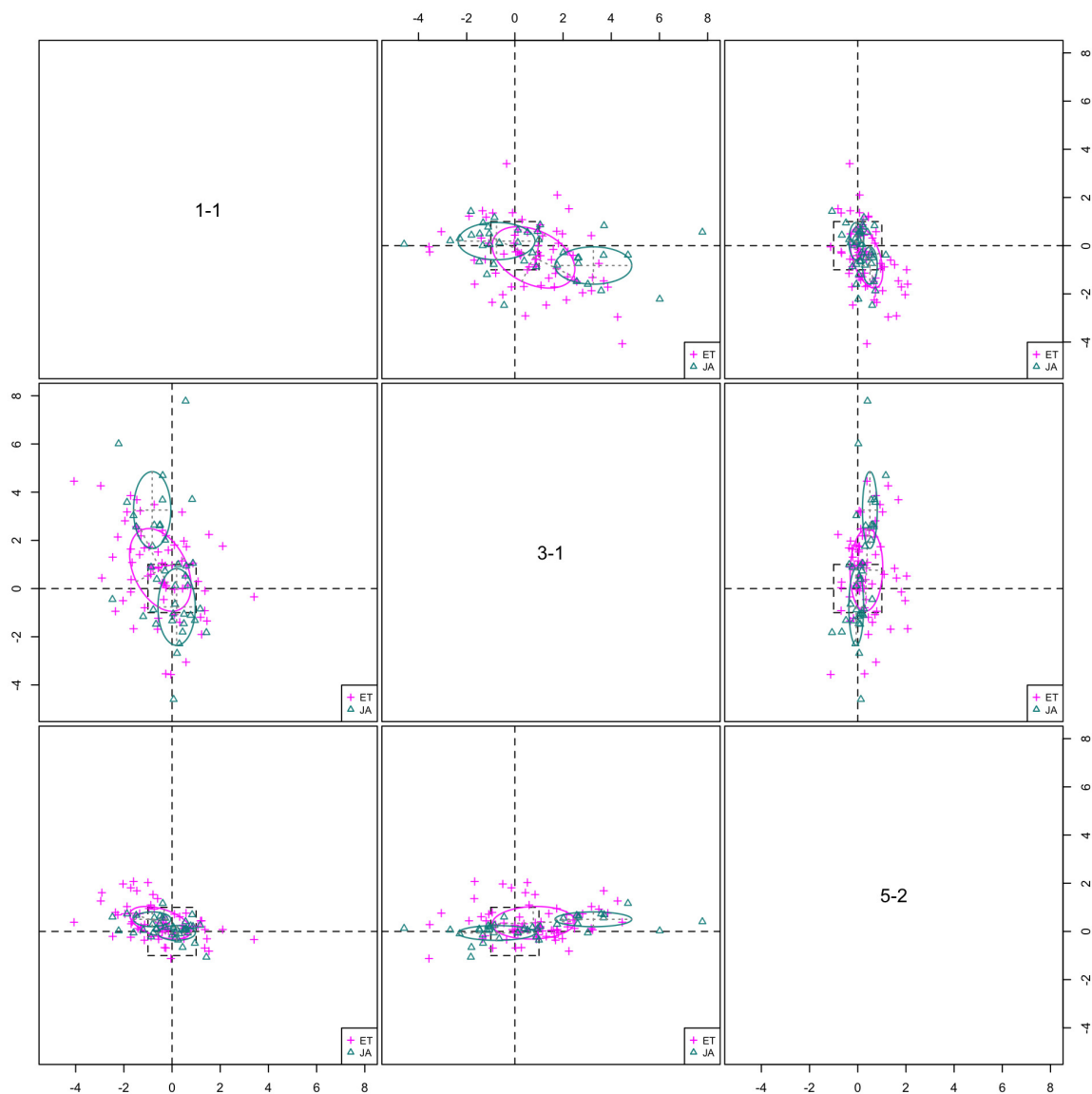
Figure 12: Scatterplot matrix for [1-1, 3-1, 5-2]
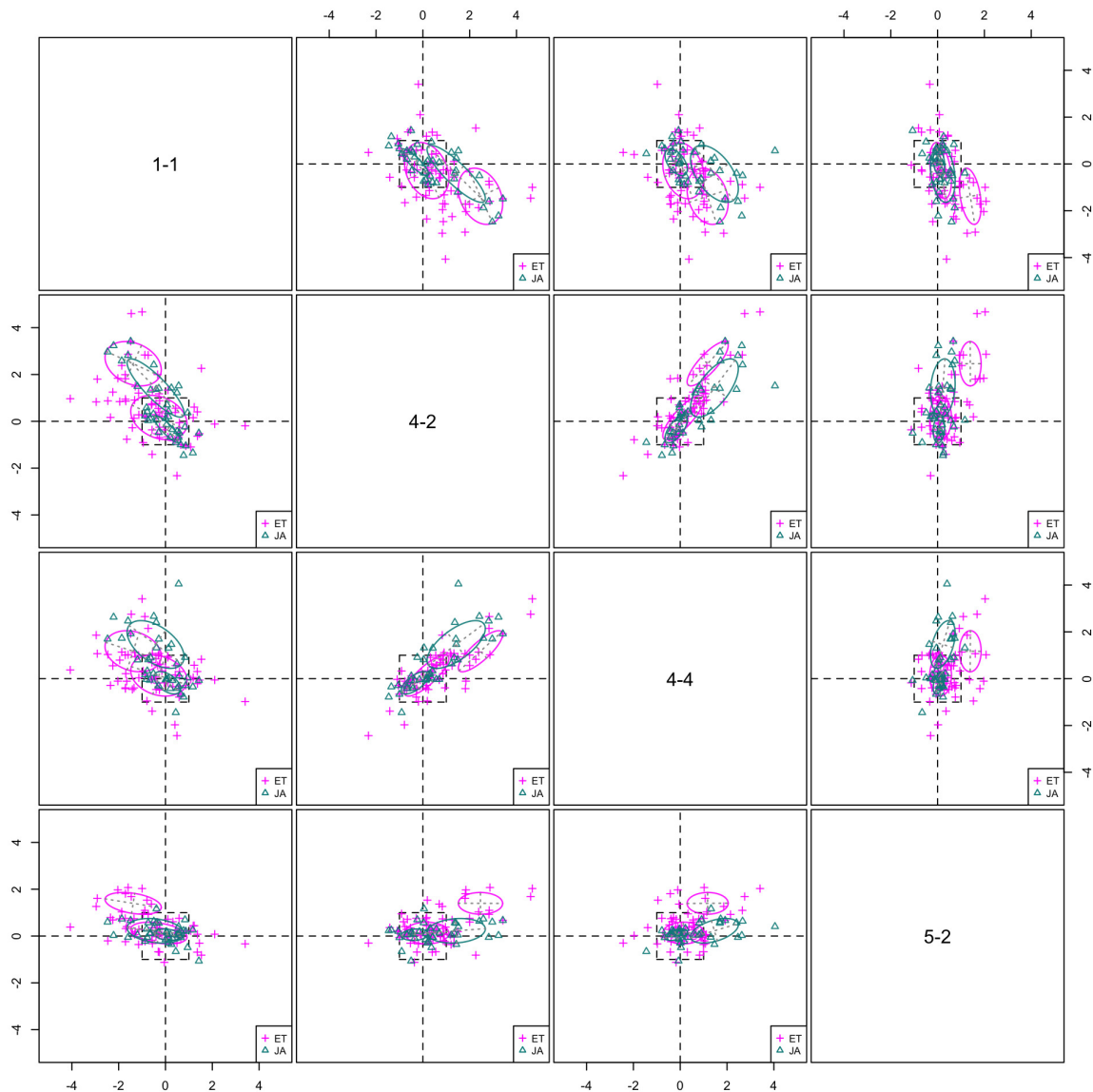
Figure 13: Scatterplot matrix for [1-1, 4-2, 4-4, 5-2]
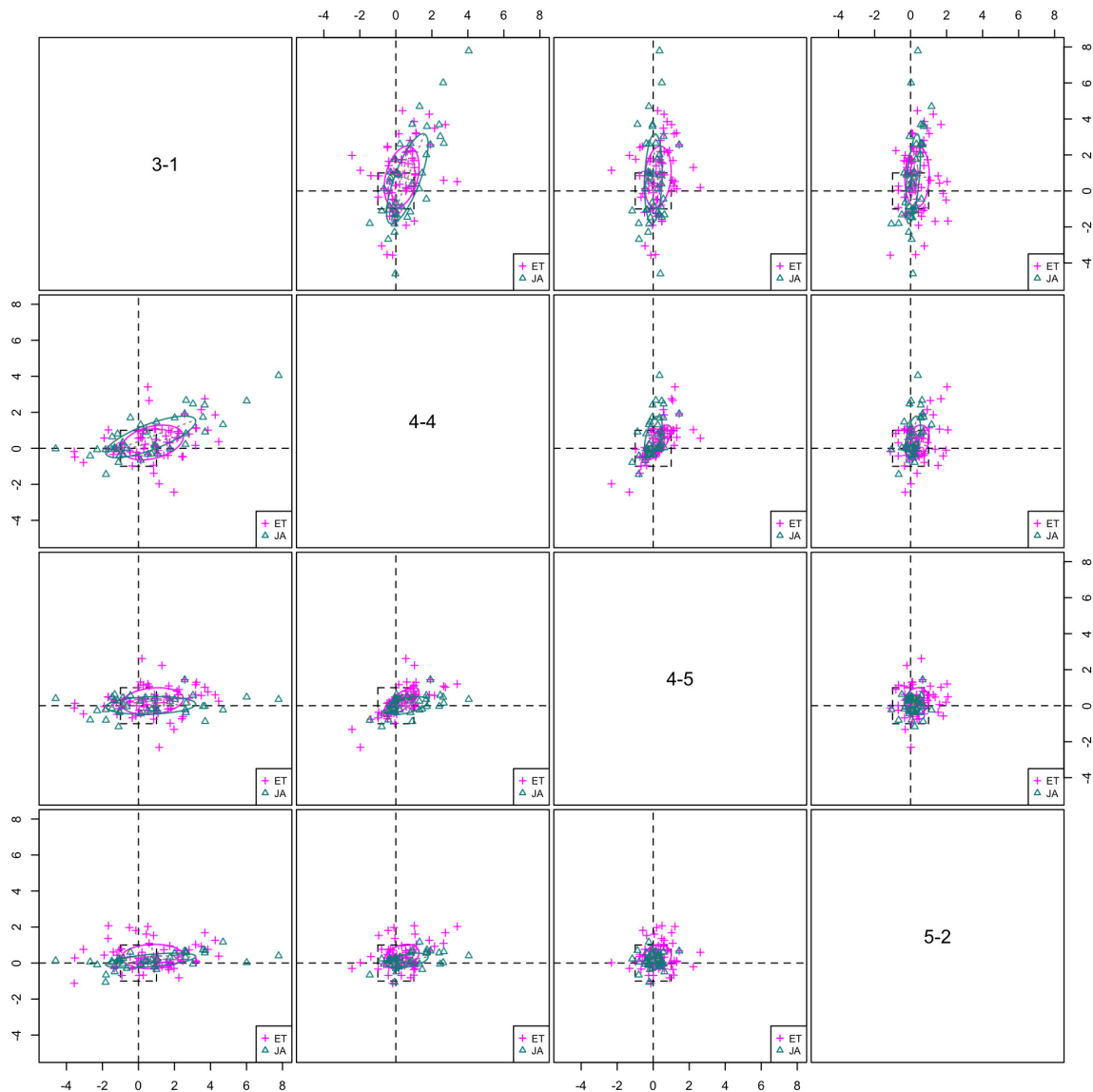
Figure 14: Scatterplot matrix for [3-1, 4-4, 4-5, 5-2]
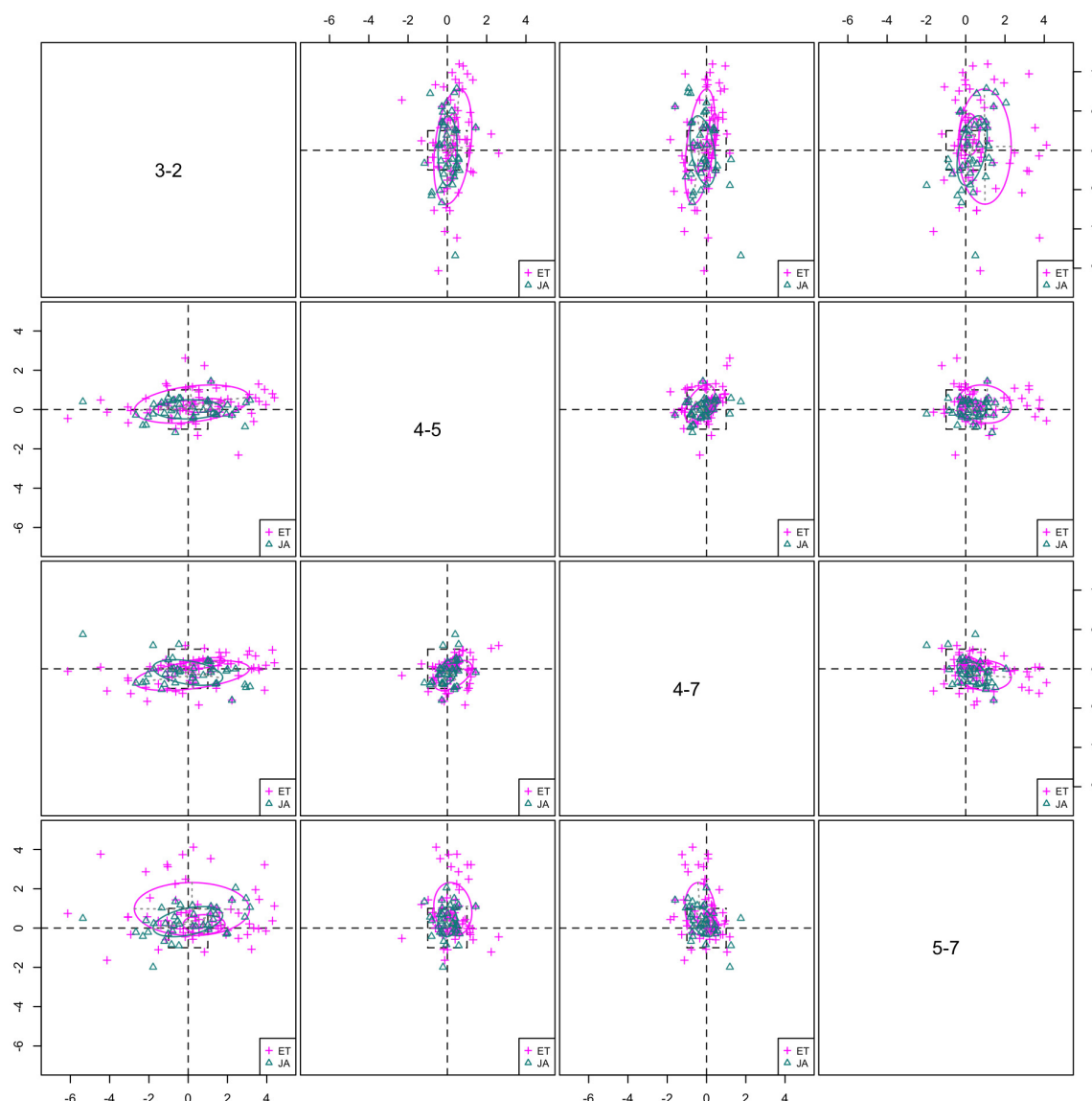
Figure 15: Scatterplot matrix for [3-2, 4-5, 4-7, 5-7]

C. BIASEDNESS AND VARIABILITY OF GSI PRODUCED BY REPEATED CV FOR LARGE SAMPLES

601 The main purpose of CV is to prevent overfitting and to correct the biased estimation of

602 misclassification error rate, while repetition of CV is performed to avoid any potentially

603 "bad" partitioning of the data in CV process.

604   To test the necessity of CV and repeated CV for large sample size, we performed

605 simulation studies to compare GSI values produced with and without CV/repeated CV

606 for large, simulated data. Steps of the simulation are as follows:

607   1. Choose a feature pair $[i, j]$, and subset the data accordingly.

608   2. Fit MclustDA model to obtain parameter estimates for each class $\hat{\Theta}_1$ and $\hat{\Theta}_2$. Denote

609   sample size in each class as N_1 and N_2.

610   3. For each class, simulate multivariate normal or mixture of normal data using $\hat{\Theta}_1$

611   and $\hat{\Theta}_2$ as model parameters.

612   − Simulate with sample sizes mN_1 and mN_2, with $m = 1, 2, 5, 8, 10$

613   4. For each simulated dataset and the original dataset, calculate GSI, first using 10×10

614   repeated CV and recording all 10 CV outputs, and then without using any CV.

615   5. Repeat steps $1 \sim 4$ for all feature pairs.

C.1   Test variability of cross-validation results for large sample size

616 As observed in Figure 16, as sample size of the simulated data increases, CV results

617 become more stable between repetitions.
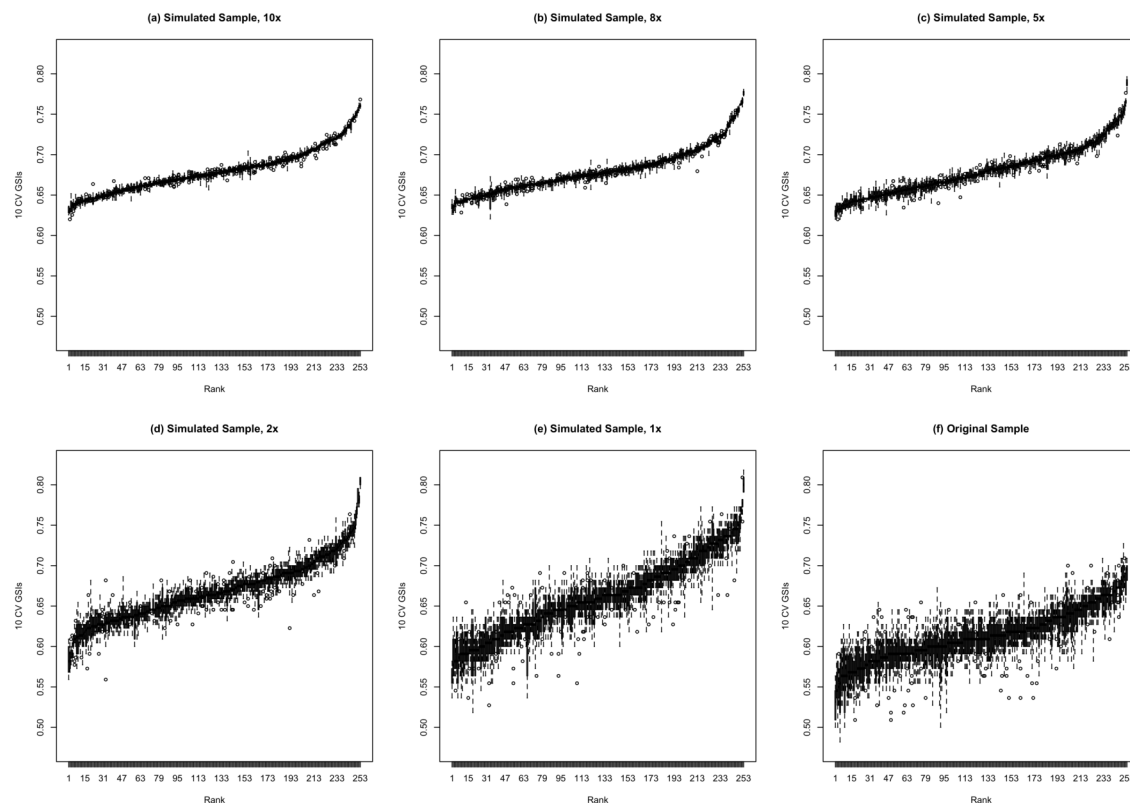
Figure 16: Variability of GSI values produced by repeated CV for simulated and original data. Feature pairs ranked by median GSI. All plots on the same scale.

C.2 Test necessity of CV for large dataset in terms of bias

As observed in Figure 17, for relatively small samples, GSI calculated without any cross-validation is often over-optimistic. The biasedness becomes ameliorated as sample size increases.
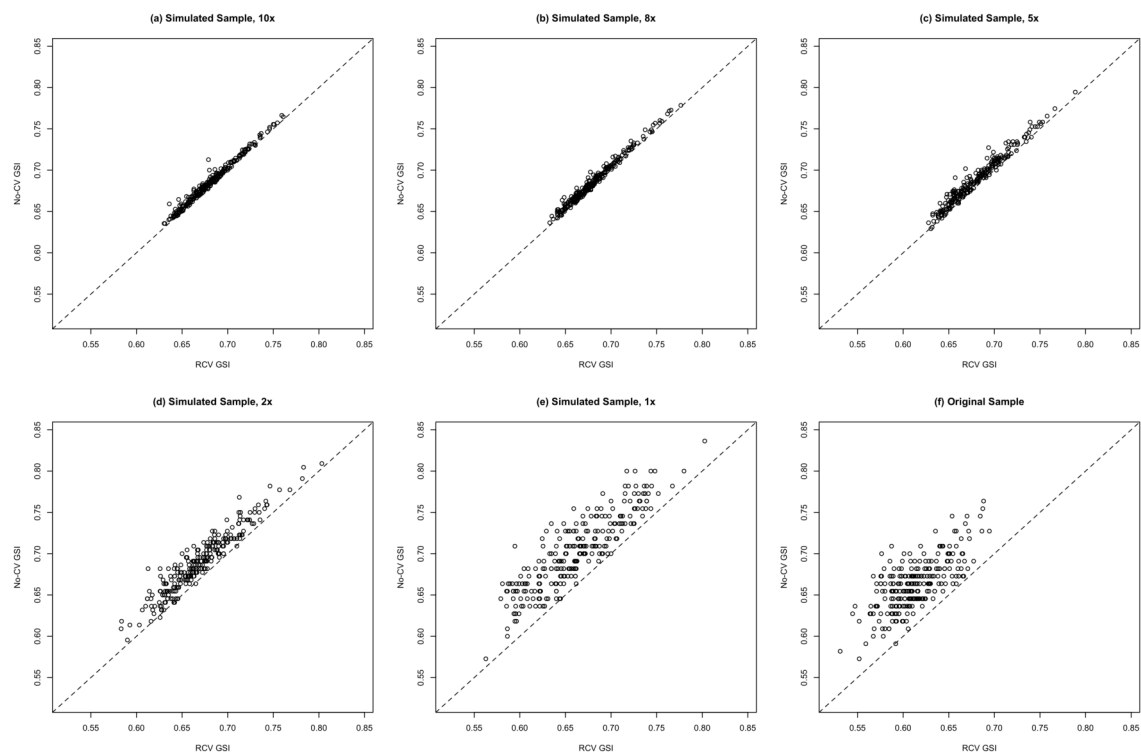
Figure 17: Comparision of GSI values produced with and without using RCV. All plots on the same scale.