The first principal component dominates in training

Kevin R. Coombes, Jing Wang, and Keith A. Baggerly

13 March 2007

1 Load the data

Start with the individual data from Novartis and the data on 10 drugs from the DTP.

```
> prep <- file.path("Tangled", "prepareData.R")
> Stangle(file.path("RNowebSource", "prepareData.Rnw"),
+     output = prep)

Writing to file Tangled/prepareData.R

> source(prep)
> rm(prep)
```

Use this function to load cached data, if it exists, and produce it from scratch otherwise.

```
> getCached <- function(rda, r) {
+     rfile <- file.path("Tangled", paste(r, "R", sep = "."))
+     rdafile <- file.path("RDataObjects", paste(rda, "Rda",
+         sep = "."))
+     if (file.exists(rdafile)) {
+         cat("loading from cache\n")
+         load(rdafile, .GlobalEnv)
+     }
+     else {
+         Stangle(file.path("RNowebSource", paste(r, "Rnw",
+             sep = ".")), output = rfile)
+         source(rfile)
+     }
+ }
```

See how easy it is to use?

```
> getCached("chemoPredictors", "predCellLines")
```

```
loading from cache
```

```
> getCached("doceGI50", "gi50ValuesOverlap")
```

```
loading from cache
```

```
> getCached("features", "wobblingFeatures")
```

```
loading from cache
```

2 Using the selected NCI60 cell lines as a training set, construct a model that predicts chemoresistance or chemosensitivity.

Their Description of the Method "The individual drug sensitivity and resistance data from the selected solid tumor [sic] NCI60 cell lines was then used in a supervised analysis using binary regression methodologies ... to develop models predictive of chemotherapeutic response.... Each signature summarizes its constituent genes as a single expression profile, and is here derived as the top principal component of that set of genes."

We used singular value decomposition (SVD) to perform a principal components analysis (PCA) on the selected features and the selected cell lines, using all the replicates in the Novartis data set. We used an implementation of the algorithm in the `ClassComparison` package that is part of a suite of tools for Object-Oriented Microarray and Proteomic Analysis (OOMPA) in R, which we developed and which is available from our web site (http://bioinformatics.mdanderson.org/software.html). We only consider models using up to the first 20 principal components.

```
> makePcaModel <- function(data, info, features) {
+     require(ClassDiscovery)
+     batchdata <- data[features, ]
+     spca <- SamplePCA(batchdata, info$Response)
+     pcaScores <- data.frame(spca@scores)
+     N <- ncol(pcaScores)
+     dimnames(pcaScores) <- list(colnames(batchdata),
+         paste("PC", 1:N, sep = ""))
+     pcaScores <- data.frame(Response = info$Response,
+         pcaScores)
+     temp <- pcaScores[, c(1, 3:21)]
+     allButOne <- glm(Response ~ ., data = temp, family = binomial(link = probit))
+     temp <- pcaScores[, 1:21]
+     massive <- glm(Response ~ ., data = temp, family = binomial(link = probit))
+     justOne <- glm(Response ~ PC1, data = temp, family = binomial(link = probit))
+     better <- step(massive, trace = 0)
+     list(spca = spca, allButOne = allButOne, better = better,
+         justOne = justOne, info = info, features = features)
+ }
```

We build models in five cases:

1. The cell lines we chose, with features based on Novartis A data
2. The cell lines we chose, with features based on averaged Novartis data
3. The cell lines Potti chose, with features based on Novartis A data
4. The cell lines Potti chose, with features based on averaged Novartis data
5. The cell lines Potti chose, with features reported by Potti in their online supplement, corrected for the off-by-one error.

```
> ourModelA <- makePcaModel(ourSubset$NL, ourSubset$Info,
+     ourStuff$Features$A)
> ourModelAvg <- makePcaModel(ourSubset$NL, ourSubset$Info,
+     ourStuff$Features$Average)
> pottiModelA <- makePcaModel(pottiSubset$NL, pottiSubset$Info,
+     pottiStuff$Features$A)
> pottiModelAvg <- makePcaModel(pottiSubset$NL, pottiSubset$Info,
+     pottiStuff$Features$Average)
> pottiModelRpt <- makePcaModel(pottiSubset$NL, pottiSubset$Info,
+     pottiOBOE)
```

3 Results

3.1 Our cell lines, features from Novartis A

```
> require(colorspace)
```

```
[1] TRUE
```

```
> attach(ourModelA)
```

The first two principal components are plotted in Figure 1, colored to indicate sensitivity or resistance. The first principal component by itself is able to separate completely the resistant and sensitive lines. This finding is consistent with the description of the methods in the paper by Potti. The same samples are plotted in Figure 2, where different colors are used to indicate the triplicates detrived from a single cell line.

We built a binary probit prediction model using all components *except* the first to try to predict sensitivity in the selected NCI60 cell lines. None of the components appeared to be significant:

```
> anova(allButOne, test = "Chi")
```

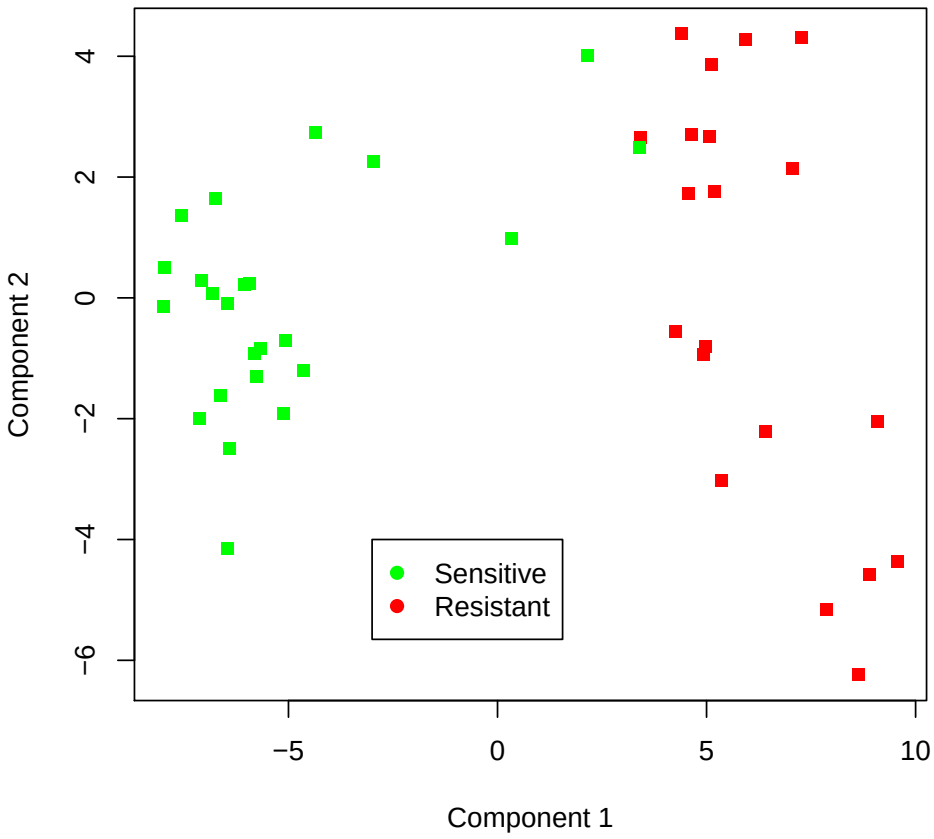

Figure 1: Plot of the sensitive (green) and resistant (red) NCI60 cell lines in the space of the first two principal components from the set of 50 differentially expressed genes.

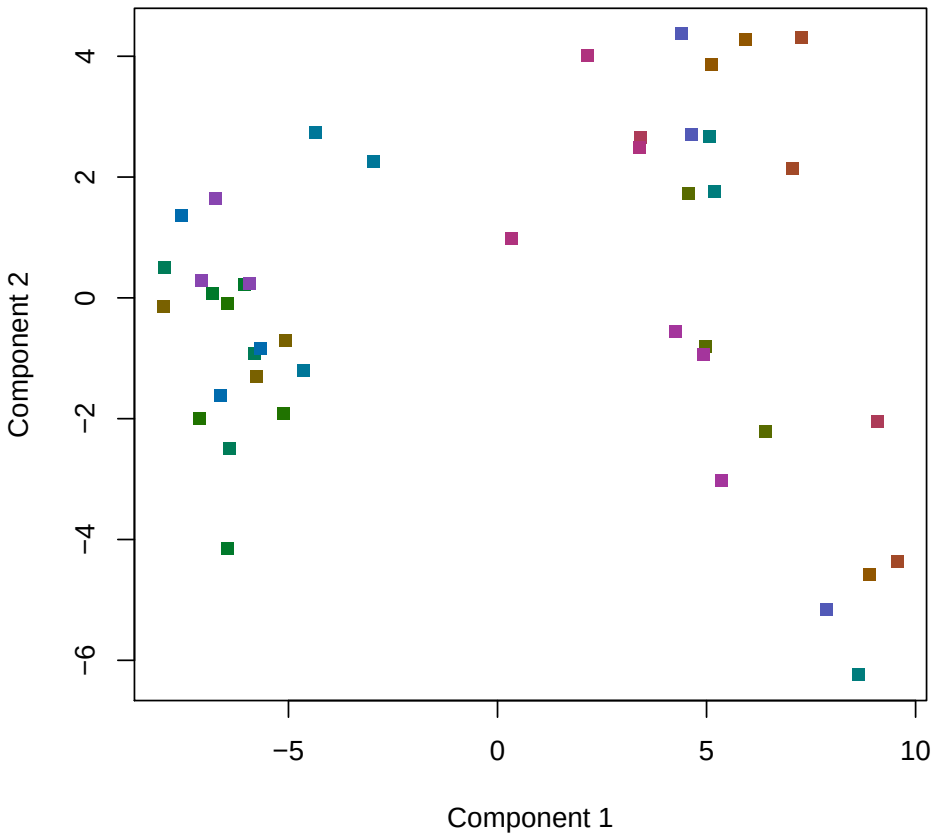

Figure 2: Plot of the first two principal components from the set of 50 differentially expressed genes, using different colors to mark replicate arrays.

```
Analysis of Deviance Table

Model: binomial, link: probit

Response: Response

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    43     60.633
PC2   1    0.004        42     60.629     0.951
PC3   1    0.713        41     59.916     0.398
PC4   1    0.309        40     59.606     0.578
PC5   1    0.098        39     59.508     0.754
PC6   1    0.261        38     59.247     0.610
PC7   1    0.029        37     59.218     0.865
PC8   1    0.144        36     59.075     0.704
PC9   1    0.041        35     59.033     0.839
PC10  1    0.807        34     58.226     0.369
PC11  1    1.527        33     56.699     0.217
PC12  1    0.262        32     56.438     0.609
PC13  1    0.100        31     56.337     0.751
PC14  1    0.867        30     55.470     0.352
PC15  1    0.035        29     55.436     0.852
PC16  1    0.445        28     54.991     0.505
PC17  1    0.277        27     54.714     0.599
PC18  1    0.105        26     54.609     0.746
PC19  1    0.039        25     54.570     0.843
PC20  1    0.030        24     54.540     0.864
```

We then built another predictive model, including the first principal component, and using the Akaike Information Criterion (AIC) in a step-wise procedure to select the best model incorporating multiple principal components.

```
> anova(better, test = "Chi")

Analysis of Deviance Table

Model: binomial, link: probit

Response: Response
```

```
Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    43     60.633
PC1   1   60.633        42  3.823e-06 6.878e-15
```

In this case, we only keep the first PC.

3.2 Our cell lines, average features

```
> detach("ourModelA")
> attach(ourModelAvg)
```

The first two principal components are plotted in Figure **??**, colored to indicate sensitivity or resistance. The first principal component by itself is able to separate completely the resistant and sensitive lines. This finding is consistent with the description of the methods in the paper by Potti. The same samples are plotted in Figure 4, where different colors are used to indicate the triplicates detrived from a single cell line.

We built a binary probit prediction model using all components *except* the first to try to predict sensitivity in the selected NCI60 cell lines. None of the components appeared to be significant:

```
> anova(allButOne, test = "Chi")

Analysis of Deviance Table

Model: binomial, link: probit

Response: Response

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    43     60.633
PC2   1    0.005        42     60.628     0.946
PC3   1    0.059        41     60.569     0.808
PC4   1    0.136        40     60.434     0.713
PC5   1    1.045        39     59.389     0.307
PC6   1    0.005        38     59.384     0.943
PC7   1    0.026        37     59.358     0.871
PC8   1    0.113        36     59.245     0.737
PC9   1    0.002        35     59.243     0.968
PC10  1    0.051        34     59.192     0.822
```

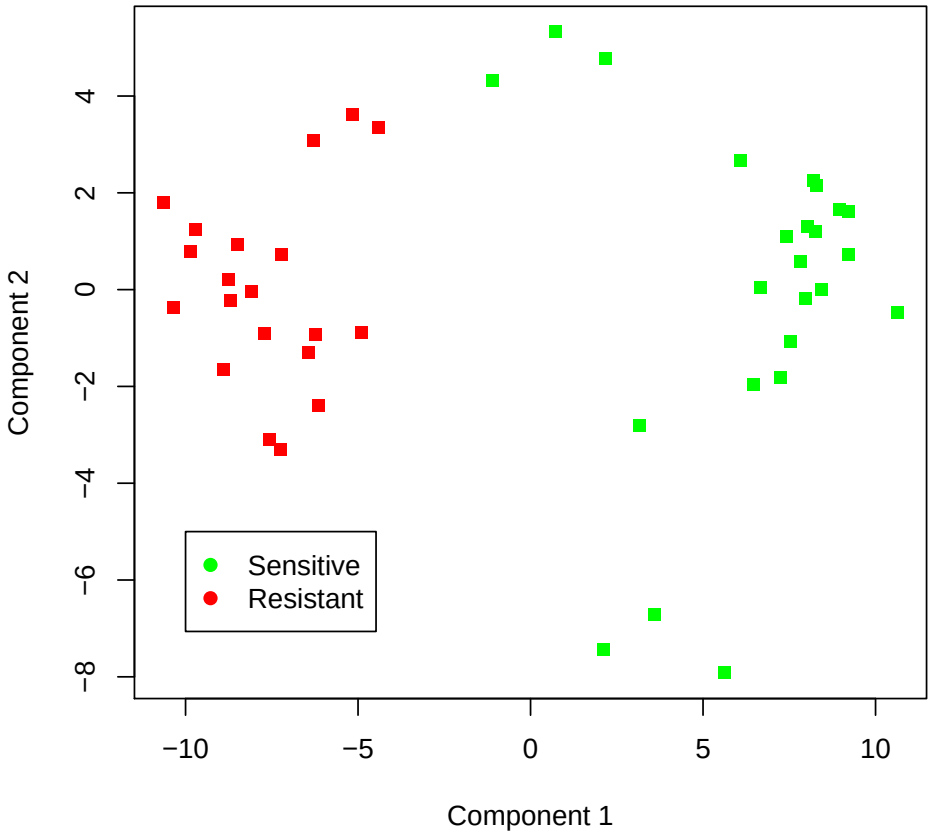

Figure 3: Plot of the sensitive (green) and resistant (red) NCI60 cell lines in the space of the first two principal components from the set of 50 differentially expressed genes.

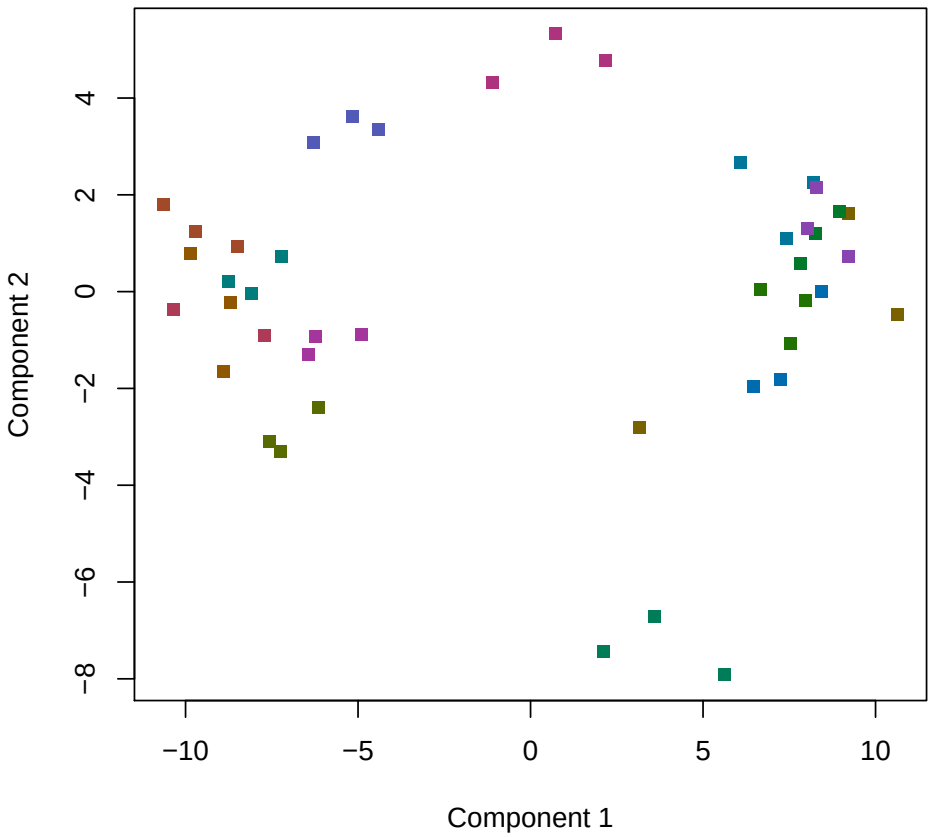

Figure 4: Plot of the first two principal components from the set of 50 differentially expressed genes, using different colors to mark replicate arrays.

```
PC11  1    0.249        33     58.943     0.618
PC12  1    0.267        32     58.677     0.606
PC13  1    0.009        31     58.667     0.924
PC14  1    0.039        30     58.628     0.843
PC15  1    0.024        29     58.605     0.878
PC16  1    0.002        28     58.603     0.967
PC17  1    0.275        27     58.328     0.600
PC18  1    0.030        26     58.298     0.863
PC19  1    0.044        25     58.254     0.834
PC20  1    0.018        24     58.236     0.894
```

We then built another predictive model, including the first principal component, and using the Akaike Information Criterion (AIC) in a step-wise procedure to select the best model incorporating multiple principal components.

```
> anova(better, test = "Chi")

Analysis of Deviance Table

Model: binomial, link: probit

Response: Response

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    43     60.633
PC1   1   60.633        42  2.525e-09 6.878e-15
```

Again, we only keep the first PC.

3.3 Potti cell lines, features from Novartis A

```
> detach("ourModelAvg")
> attach(pottiModelA)
```

The first two principal components are plotted in Figure 5, colored to indicate sensitivity or resistance. The first principal component by itself is able to separate completely the resistant and sensitive lines. This finding is consistent with the description of the methods in the paper by Potti. The same samples are plotted in Figure 6, where different colors are used to indicate the triplicates detrived from a single cell line.

We built a binary probit prediction model using all components *except* the first to try to predict sensitivity in the selected NCI60 cell lines. None of the components appeared to be significant:

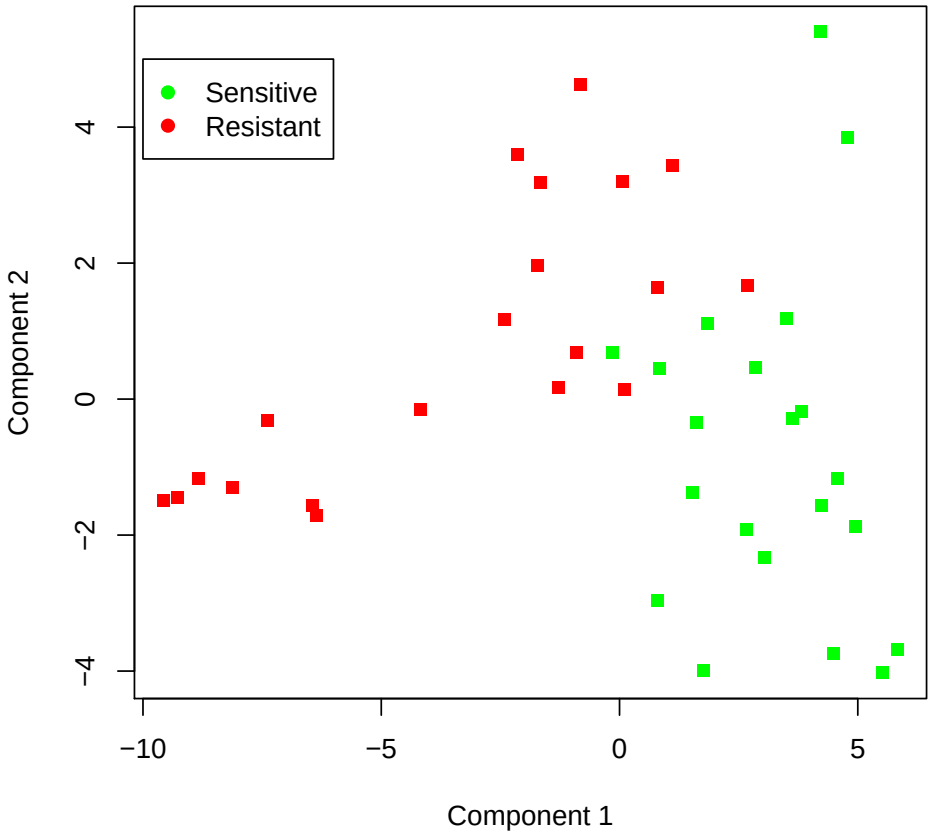

Figure 5: Plot of the sensitive (green) and resistant (red) NCI60 cell lines in the space of the first two principal components from the set of 50 differentially expressed genes.

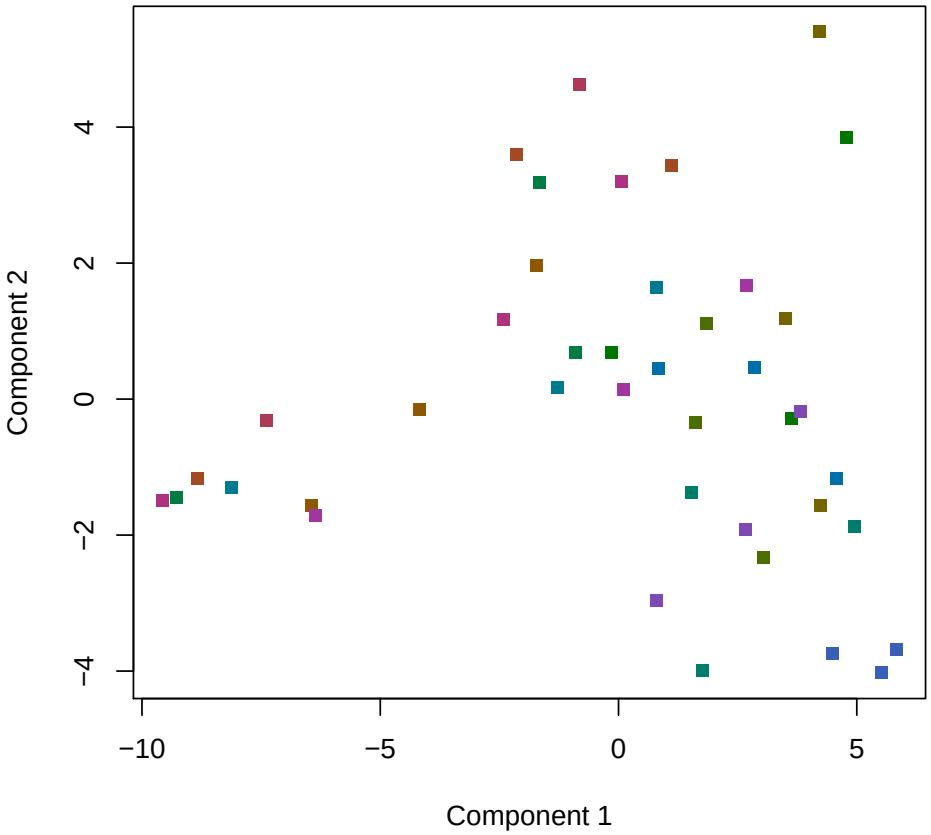

Figure 6: Plot of the first two principal components from the set of 50 differentially expressed genes, using different colors to mark replicate arrays.

```
> anova(allButOne, test = "Chi")

Analysis of Deviance Table

Model: binomial, link: probit

Response: Response

Terms added sequentially (first to last)


     Df  Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    40     56.814
PC2   1     4.926        39     51.887     0.026
PC3   1     0.410        38     51.477     0.522
PC4   1     3.069        37     48.408     0.080
PC5   1     0.211        36     48.198     0.646
PC6   1 7.428e-05        35     48.198     0.993
PC7   1     0.123        34     48.075     0.726
PC8   1     0.073        33     48.002     0.786
PC9   1 4.261e-04        32     48.001     0.984
PC10  1     0.429        31     47.572     0.512
PC11  1     0.899        30     46.673     0.343
PC12  1     2.784        29     43.889     0.095
PC13  1     0.112        28     43.776     0.737
PC14  1     0.321        27     43.455     0.571
PC15  1     0.576        26     42.879     0.448
PC16  1     0.226        25     42.653     0.635
PC17  1     0.248        24     42.405     0.618
PC18  1     1.571        23     40.834     0.210
PC19  1     3.987        22     36.847     0.046
PC20  1     0.071        21     36.775     0.790
```

We then built another predictive model, including the first principal component, and using the Akaike Information Criterion (AIC) in a step-wise procedure to select the best model incorporating multiple principal components.

```
> anova(better, test = "Chi")

Analysis of Deviance Table

Model: binomial, link: probit
```

```
Response: Response

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    40     56.814
PC1   1   37.968        39     18.846 7.192e-10
PC2   1    6.404        38     12.441     0.011
PC4   1   12.441        37  1.841e-08 4.199e-04
```

Here we keep three principal components: 1, 2, and 4.

3.4 Potti cell lines, average features

```
> detach("pottiModelA")
> attach(pottiModelAvg)
```

The first two principal components are plotted in Figure 7, colored to indicate sensitivity or resistance. The first principal component by itself is able to separate completely the resistant and sensitive lines. This finding is consistent with the description of the methods in the paper by Potti. The same samples are plotted in Figure 8, where different colors are used to indicate the triplicates detrived from a single cell line.

We built a binary probit prediction model using all components *except* the first to try to predict sensitivity in the selected NCI60 cell lines. None of the components appeared to be significant:

```
> anova(allButOne, test = "Chi")

Analysis of Deviance Table

Model: binomial, link: probit

Response: Response

Terms added sequentially (first to last)


     Df  Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    40     56.814
PC2   1     0.461        39     56.352     0.497
PC3   1     0.265        38     56.087     0.606
PC4   1     0.150        37     55.937     0.698
PC5   1     0.106        36     55.831     0.745
PC6   1     0.043        35     55.788     0.836
```

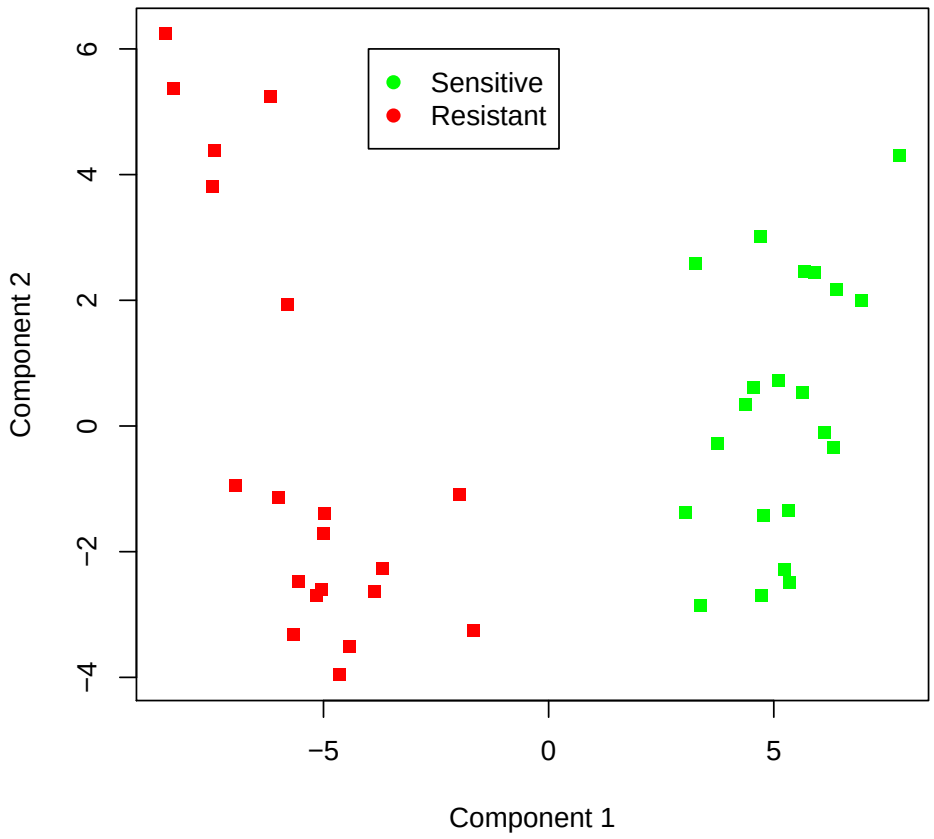

Figure 7: Plot of the sensitive (green) and resistant (red) NCI60 cell lines in the space of the first two principal components from the set of 50 differentially expressed genes.

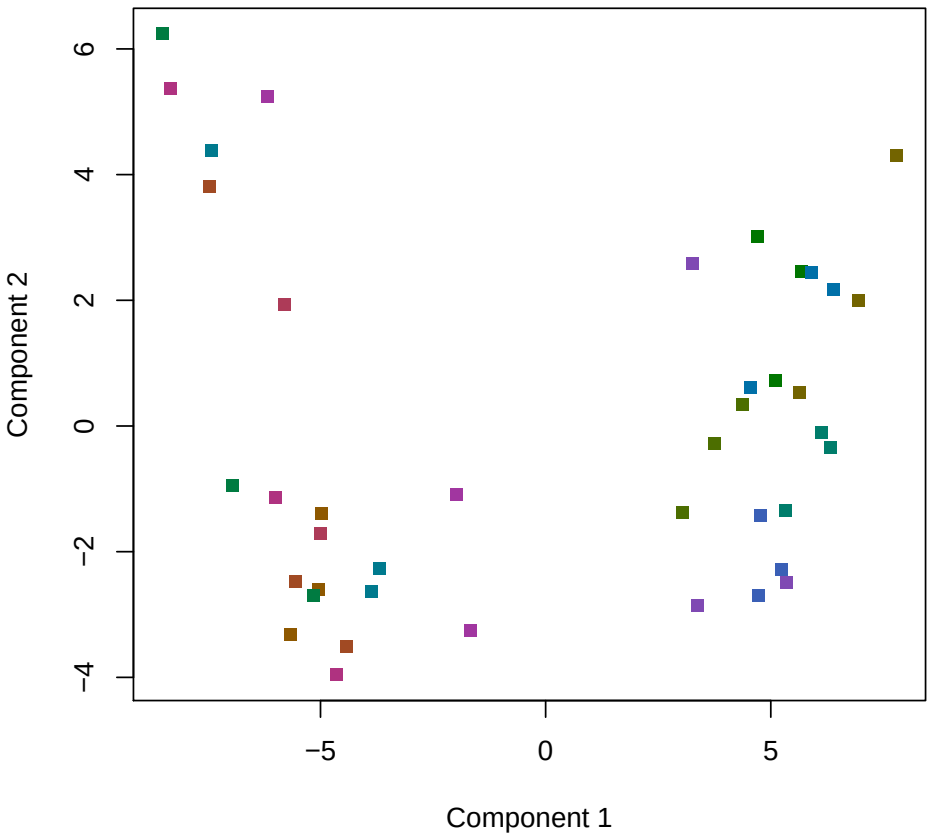

Figure 8: Plot of the first two principal components from the set of 50 differentially expressed genes, using different colors to mark replicate arrays.

```
PC7   1     0.052        34     55.736     0.820
PC8   1 1.447e-06        33     55.736     0.999
PC9   1     0.307        32     55.429     0.580
PC10  1     0.008        31     55.421     0.928
PC11  1     0.019        30     55.402     0.891
PC12  1     0.420        29     54.982     0.517
PC13  1     0.011        28     54.970     0.915
PC14  1     0.003        27     54.967     0.953
PC15  1     0.039        26     54.928     0.843
PC16  1     0.002        25     54.925     0.961
PC17  1     0.256        24     54.669     0.613
PC18  1     0.001        23     54.668     0.976
PC19  1     0.017        22     54.652     0.897
PC20  1     0.003        21     54.648     0.955
```

We then built another predictive model, including the first principal component, and using the Akaike Information Criterion (AIC) in a step-wise procedure to select the best model incorporating multiple principal components.

```
> anova(better, test = "Chi")

Analysis of Deviance Table

Model: binomial, link: probit

Response: Response

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    40     56.814
PC1   1   56.814        39  6.753e-10 4.791e-14
```

In this case, we only keep the first PC.

3.5 Potti cell lines, reported features

```
> detach("pottiModelAvg")
> attach(pottiModelRpt)
```

The first two principal components are plotted in Figure 9, colored to indicate sensitivity or resistance. The first principal component by itself is able to separate completely the resistant and sensitive lines. This finding is consistent with the description of the methods in the paper by Potti.

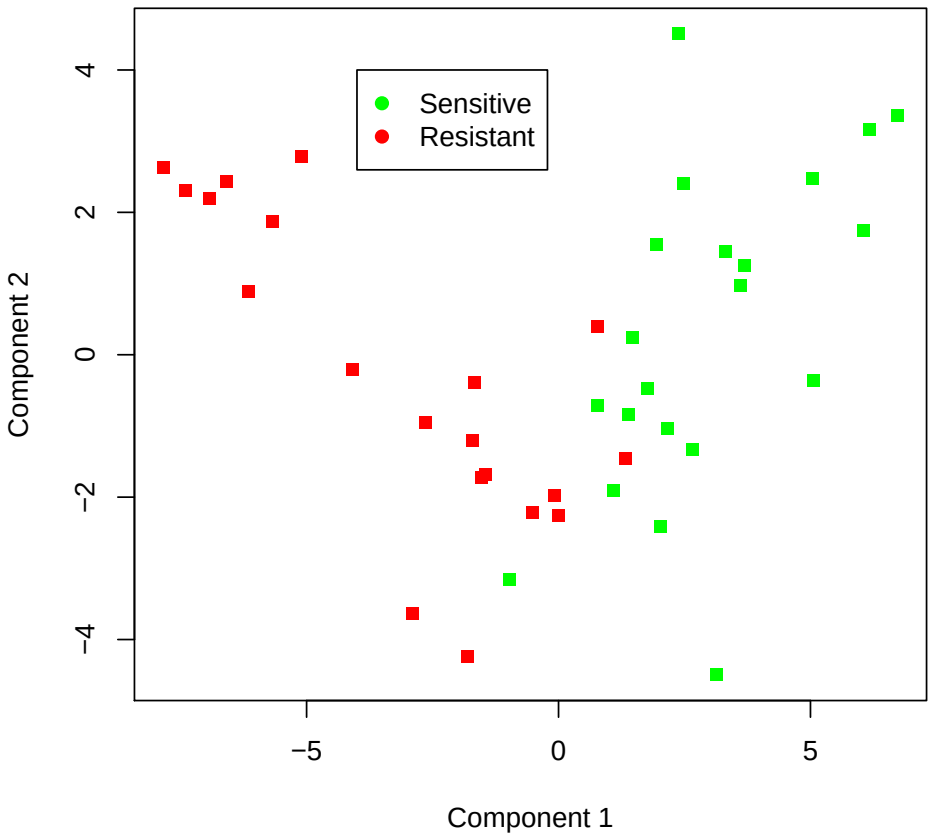

Figure 9: Plot of the sensitive (green) and resistant (red) NCI60 cell lines in the space of the first two principal components from the set of 50 differentially expressed genes.

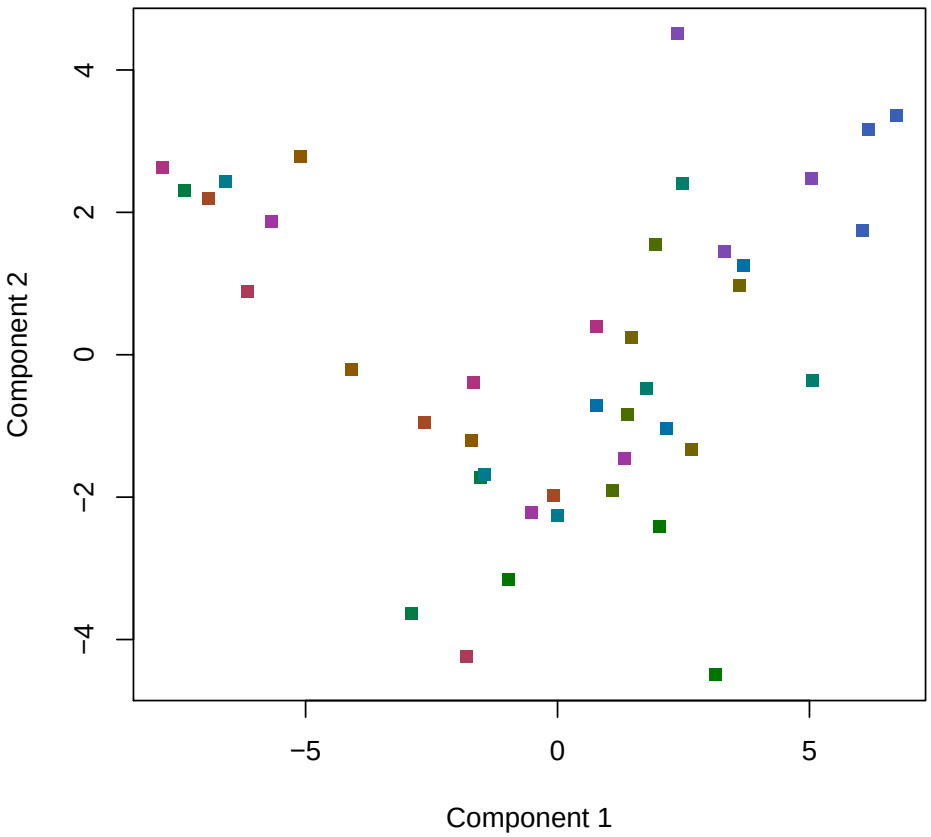

Figure 10: Plot of the first two principal components from the set of 50 differentially expressed genes, using different colors to mark replicate arrays.

The same samples are plotted in Figure 10, where different colors are used to indicate the triplicates detrived from a single cell line.

We built a binary probit prediction model using all components *except* the first to try to predict sensitivity in the selected NCI60 cell lines. None of the components appeared to be significant:

```
> anova(allButOne, test = "Chi")

Analysis of Deviance Table

Model: binomial, link: probit

Response: Response

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    40     56.814
PC2   1    0.833        39     55.981     0.361
PC3   1    0.089        38     55.892     0.766
PC4   1    3.341        37     52.551     0.068
PC5   1    0.645        36     51.907     0.422
PC6   1    0.046        35     51.861     0.829
PC7   1    3.743        34     48.117     0.053
PC8   1    0.304        33     47.814     0.582
PC9   1    0.001        32     47.813     0.975
PC10  1    0.089        31     47.724     0.765
PC11  1    0.217        30     47.507     0.642
PC12  1    0.063        29     47.444     0.801
PC13  1    0.560        28     46.884     0.454
PC14  1    0.592        27     46.292     0.442
PC15  1    1.818        26     44.473     0.177
PC16  1    0.021        25     44.452     0.884
PC17  1    0.161        24     44.291     0.688
PC18  1    0.086        23     44.205     0.769
PC19  1    0.175        22     44.030     0.676
PC20  1    0.298        21     43.732     0.585
```

We then built another predictive model, including the first principal component, and using the Akaike Information Criterion (AIC) in a step-wise procedure to select the best model incorporating multiple principal components.

```
> anova(better, test = "Chi")
```

```
Analysis of Deviance Table

Model: binomial, link: probit

Response: Response

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    40     56.814
PC1   1   41.422        39     15.391 1.226e-10
PC7   1    4.209        38     11.182     0.040
PC15  1   11.182        37  9.493e-09     0.001
```

Three principal components are kept in th model: 1, 7, and 15, although PC1 is dominant.

```
> detach("pottiModelRpt")
```

4 Wrapup

```
> rm(makePcaModel)
```

```
> save(pottiModelA, pottiModelAvg, pottiModelRpt, ourModelA,
+     ourModelAvg, file = file.path("RDataObjects", "pcaModels.Rda"))
```