

# PCA, Mahalanobis Distance, and Outliers

Kevin R. Coombes

4 November 2011

## Contents

|          |                       |          |
|----------|-----------------------|----------|
| <b>1</b> | <b>Simulated Data</b> | <b>1</b> |
| <b>2</b> | <b>PCA</b>            | <b>1</b> |
| <b>3</b> | <b>A Second Round</b> | <b>5</b> |
| <b>4</b> | <b>A Final Round</b>  | <b>8</b> |
| <b>5</b> | <b>Appendix</b>       | <b>8</b> |

## 1 Simulated Data

We simulate a dataset.

```
> set.seed(564684)
> nSamples <- 30
> nGenes <- 3000
> dataset <- matrix(rnorm(nSamples*nGenes), ncol=nSamples, nrow=nGenes)
> dimnames(dataset) <- list(paste("G", 1:nGenes, sep=''),
+                             paste("S", 1:nSamples, sep=''))
```

Now we make two of the entries into distinct outliers.

```
> nShift <- 300
> affected <- sample(nGenes, nShift)
> dataset[affected,1] <- dataset[affected,1] + rnorm(nShift, 1, 1)
> dataset[affected,2] <- dataset[affected,2] + rnorm(nShift, 1, 1)
```

## 2 PCA

We start with a principal components analysis (PCA) of this dataset. A plot of the samples against the first two principal components (PCs) shows two very clear outliers (**Figure 1**).

```
> library(ClassDiscovery)
> spca <- SamplePCA(dataset)
```

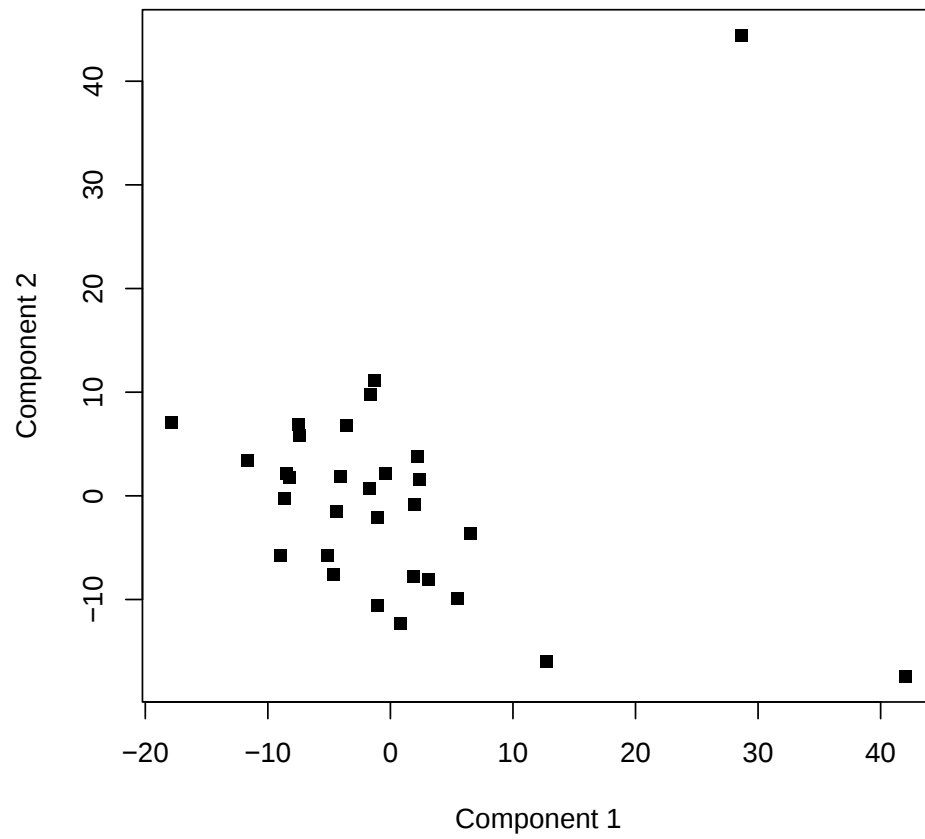


Figure 1: Principal components plot of the samples.

We want to explore the possibility of an outlier more formally. First, we look at the cumulative amount of variance explained by the PCs:

```
> round(cumsum(spca@variances)/sum(spca@variances), digits=2)

[1] 0.04 0.08 0.12 0.16 0.20 0.24 0.28 0.31 0.35 0.39 0.42 0.46 0.49 0.53 0.56 0.60
[17] 0.63 0.66 0.69 0.73 0.76 0.79 0.82 0.85 0.88 0.91 0.94 0.97 1.00 1.00
```

We see that we need 20 components in order to explain 70% of the variation in the data. Next, we compute the Mahalanobis distance of each sample from the center of an  $N$ -dimensional principal component space. We apply the `mahalanobisQC` function using different numbers of components between 2 and 20.

```
> maha2 <- mahalanobisQC(spca, 2)
> maha5 <- mahalanobisQC(spca, 5)
> maha10 <- mahalanobisQC(spca, 10)
> maha20 <- mahalanobisQC(spca, 20)
> myd <- data.frame(maha2, maha5, maha10, maha20)
> colnames(myd) <- paste("N", rep(c(2, 5, 10, 20), each=2),
+                          rep(c(".statistic", ".p.value"), 4), sep='')
```

The theory says that, under the null hypothesis that all samples arise from the same multivariate normal distribution, the distance from the center of a  $d$ -dimensional PC space should follow a chi-squared distribution with  $d$  degrees of freedom. This theory lets us compute  $p$ -values associated with the Mahalanobis distances for each sample (**Table 1**). We see that the samples S1 and S2 are outliers, at least when we look at the first 2, 5, or, 10 components. However, sample S2 is not quite significant (at the 5% level) when we get out to 20 components. This can occur when there are multiple outliers because of the “inflated” variance estimates coming from the outliers themselves.

|     | N2.statistic | N2.p.value | N5.statistic | N5.p.value | N10.statistic | N10.p.value | N20.statistic | N20.p.value |
|-----|--------------|------------|--------------|------------|---------------|-------------|---------------|-------------|
| S1  | 27.4         | 0.0000     | 32.6         | 0.0000     | 34.1          | 0.0002      | 38.3          | 0.0083      |
| S2  | 43.1         | 0.0000     | 43.4         | 0.0000     | 44.7          | 0.0000      | 46.8          | 0.0006      |
| S3  | 0.2          | 0.9269     | 0.7          | 0.9850     | 8.7           | 0.5634      | 22.5          | 0.3161      |
| S4  | 0.4          | 0.8090     | 2.0          | 0.8460     | 3.4           | 0.9711      | 18.0          | 0.5862      |
| S5  | 0.5          | 0.7778     | 0.6          | 0.9860     | 1.9           | 0.9969      | 31.3          | 0.0510      |
| S6  | 0.5          | 0.7969     | 7.2          | 0.2076     | 12.6          | 0.2485      | 20.4          | 0.4332      |
| S7  | 3.0          | 0.2281     | 8.7          | 0.1197     | 17.9          | 0.0558      | 26.5          | 0.1502      |
| S8  | 1.0          | 0.6038     | 2.5          | 0.7801     | 11.7          | 0.3036      | 20.0          | 0.4606      |
| S9  | 0.6          | 0.7563     | 1.5          | 0.9157     | 9.7           | 0.4643      | 24.7          | 0.2132      |
| S10 | 0.0          | 0.9810     | 9.1          | 0.1067     | 12.9          | 0.2277      | 25.3          | 0.1886      |
| S11 | 0.5          | 0.7672     | 3.1          | 0.6820     | 18.6          | 0.0451      | 23.6          | 0.2591      |
| S12 | 0.9          | 0.6513     | 5.7          | 0.3382     | 11.7          | 0.3078      | 20.5          | 0.4240      |
| S13 | 0.9          | 0.6337     | 1.1          | 0.9534     | 8.5           | 0.5813      | 15.8          | 0.7281      |
| S14 | 0.0          | 0.9875     | 2.5          | 0.7725     | 7.0           | 0.7277      | 17.2          | 0.6384      |
| S15 | 0.1          | 0.9706     | 0.2          | 0.9986     | 10.1          | 0.4300      | 15.2          | 0.7662      |
| S16 | 1.2          | 0.5405     | 2.7          | 0.7437     | 13.3          | 0.2065      | 25.8          | 0.1725      |
| S17 | 1.0          | 0.6042     | 5.7          | 0.3380     | 7.6           | 0.6716      | 14.7          | 0.7940      |
| S18 | 0.6          | 0.7483     | 5.3          | 0.3808     | 7.5           | 0.6770      | 17.2          | 0.6396      |
| S19 | 0.5          | 0.7950     | 5.1          | 0.4073     | 14.9          | 0.1374      | 21.7          | 0.3551      |
| S20 | 0.8          | 0.6784     | 2.1          | 0.8307     | 8.3           | 0.6016      | 21.0          | 0.3971      |
| S21 | 0.8          | 0.6727     | 10.9         | 0.0536     | 16.9          | 0.0765      | 23.8          | 0.2513      |
| S22 | 0.6          | 0.7336     | 3.5          | 0.6159     | 4.3           | 0.9331      | 13.0          | 0.8764      |
| S23 | 0.2          | 0.9276     | 1.9          | 0.8691     | 6.2           | 0.7944      | 17.9          | 0.5931      |
| S24 | 0.2          | 0.9230     | 3.8          | 0.5716     | 7.2           | 0.7095      | 22.9          | 0.2948      |
| S25 | 0.6          | 0.7484     | 8.7          | 0.1223     | 13.8          | 0.1802      | 25.4          | 0.1870      |
| S26 | 0.0          | 0.9796     | 1.7          | 0.8921     | 8.9           | 0.5437      | 19.3          | 0.5036      |
| S27 | 1.1          | 0.5698     | 6.0          | 0.3084     | 9.4           | 0.4986      | 17.0          | 0.6547      |
| S28 | 0.0          | 0.9838     | 2.8          | 0.7300     | 3.6           | 0.9638      | 24.8          | 0.2076      |
| S29 | 0.7          | 0.7123     | 1.9          | 0.8641     | 6.4           | 0.7848      | 16.7          | 0.6732      |
| S30 | 3.4          | 0.1846     | 3.5          | 0.6238     | 4.9           | 0.8951      | 18.4          | 0.5640      |

Table 1: Mahalanobis distance (with unadjusted p-values) of each sample from the center of N-dimensional principal component space.

### 3 A Second Round

Now we repeat the PCA after removing the one definite outlier. Sample S2 still stands out as “not like the others” (**Figure 2**).

```
> reduced <- dataset[,-1]
> dim(reduced)

[1] 3000    29

> spca <- SamplePCA(reduced)
> round(cumsum(spca@variances)/sum(spca@variances), digits=2)

[1] 0.04 0.09 0.13 0.17 0.21 0.24 0.28 0.32 0.36 0.40 0.43 0.47 0.51 0.54 0.58 0.61
[17] 0.65 0.68 0.71 0.75 0.78 0.81 0.85 0.88 0.91 0.94 0.97 1.00 1.00
```

And we can recompute the mahalanobis distances (**Table 2**). Here we see that even out at the level of 20 components, this sample remains an outlier.

```
> maha20 <- mahalanobisQC(spca, 20)
```

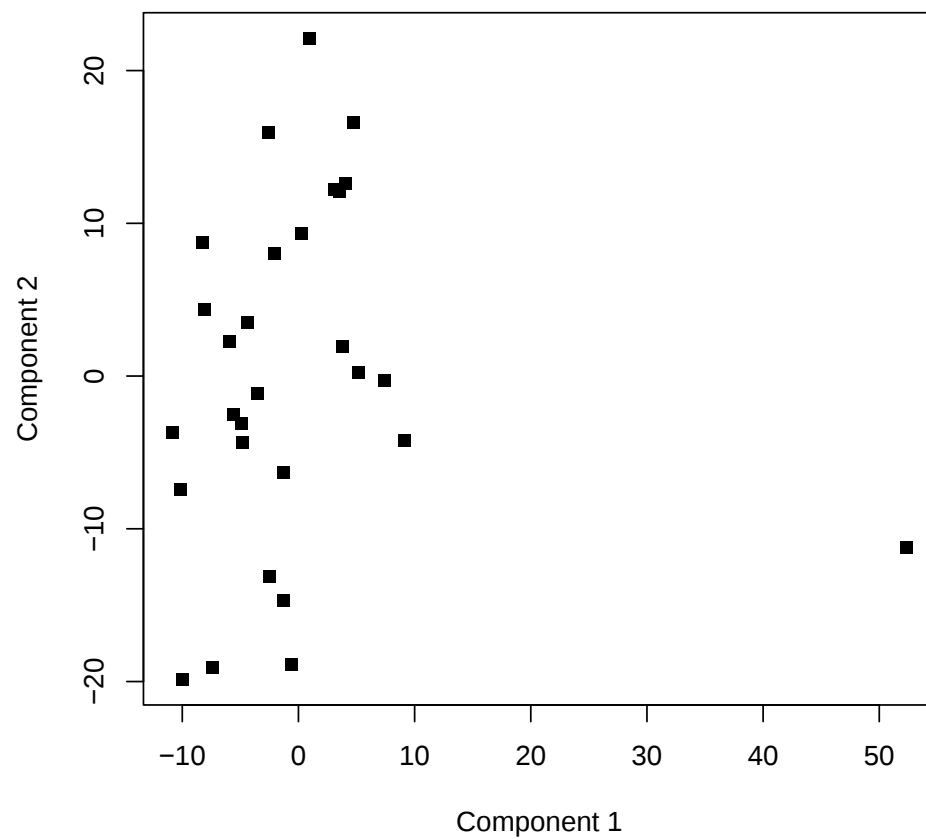


Figure 2: Principal components plot of the normal control samples, after omitting an extreme outlier.

|     | statistic | p.value |
|-----|-----------|---------|
| S2  | 94.8      | 0.0000  |
| S3  | 23.5      | 0.2670  |
| S4  | 21.5      | 0.3689  |
| S5  | 32.6      | 0.0375  |
| S6  | 20.6      | 0.4185  |
| S7  | 24.7      | 0.2130  |
| S8  | 19.5      | 0.4901  |
| S9  | 24.8      | 0.2099  |
| S10 | 23.9      | 0.2461  |
| S11 | 20.9      | 0.4019  |
| S12 | 19.3      | 0.5029  |
| S13 | 15.3      | 0.7603  |
| S14 | 17.4      | 0.6271  |
| S15 | 30.2      | 0.0673  |
| S16 | 25.8      | 0.1712  |
| S17 | 18.8      | 0.5357  |
| S18 | 17.2      | 0.6378  |
| S19 | 22.4      | 0.3214  |
| S20 | 20.3      | 0.4363  |
| S21 | 23.6      | 0.2596  |
| S22 | 11.7      | 0.9246  |
| S23 | 17.5      | 0.6188  |
| S24 | 22.0      | 0.3430  |
| S25 | 24.3      | 0.2277  |
| S26 | 18.3      | 0.5677  |
| S27 | 16.6      | 0.6769  |
| S28 | 28.1      | 0.1079  |
| S29 | 16.3      | 0.6952  |
| S30 | 22.7      | 0.3039  |

Table 2: Mahalanobis distance (with unadjusted p-values) of each sample from the center of 20-dimensional principal component space.

## 4 A Final Round

We repeat the analysis after removing one more outlier.

```
> red2 <- reduced[,-1]
> dim(red2)

[1] 3000    28

> spca <- SamplePCA(red2)
> round(cumsum(spca@variances)/sum(spca@variances), digits=2)

[1] 0.04 0.09 0.13 0.17 0.21 0.25 0.29 0.33 0.37 0.41 0.45 0.48 0.52 0.56 0.59 0.63
[17] 0.67 0.70 0.74 0.77 0.81 0.84 0.87 0.91 0.94 0.97 1.00 1.00
```

And we can recompute the mahalanobis distances (**Table 3**). At this point, there are no outliers.

```
> maha20 <- mahalanobisQC(spca, 20)
```

## 5 Appendix

This analysis was performed in the following directory:

```
> getwd()

[1] "C:/Users/KRC/AppData/Local/Temp/Rtmpwtkxwg/Rbuild9006ea85891/ClassDiscovery/vignettes"
```

This analysis was performed in the following software environment:

```
> sessionInfo()

R version 4.6.0 (2026-04-24 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 26200)

Matrix products: default
  LAPACK version 3.12.1

locale:
[1] LC_COLLATE=C                      LC_CTYPE=English_United States.utf8
[3] LC_MONETARY=English_United States.utf8 LC_NUMERIC=C
[5] LC_TIME=English_United States.utf8

time zone: America/New_York
tzcode source: internal

attached base packages:
[1] stats      graphics  grDevices  utils      datasets  methods   base

other attached packages:
```

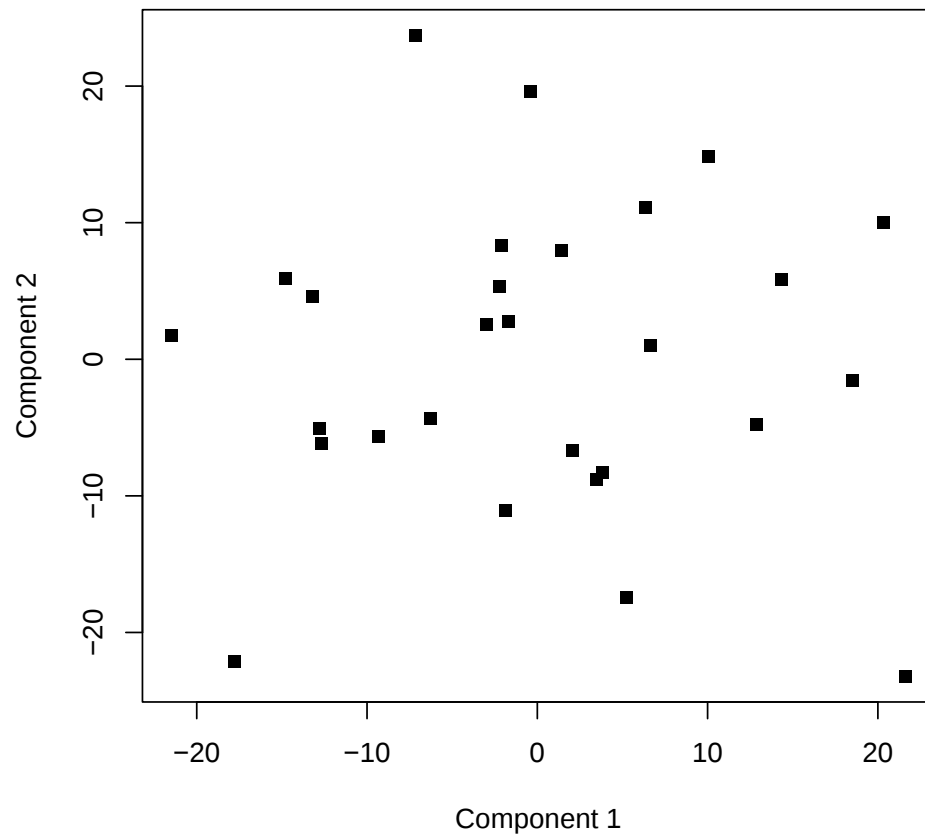


Figure 3: Principal components plot of the normal control samples, after omitting an extreme outlier.

|     | statistic | p.value |
|-----|-----------|---------|
| S3  | 25.3      | 0.1896  |
| S4  | 21.3      | 0.3773  |
| S5  | 22.9      | 0.2913  |
| S6  | 20.0      | 0.4575  |
| S7  | 23.8      | 0.2533  |
| S8  | 19.7      | 0.4789  |
| S9  | 23.5      | 0.2629  |
| S10 | 23.4      | 0.2679  |
| S11 | 21.5      | 0.3681  |
| S12 | 19.3      | 0.5043  |
| S13 | 18.4      | 0.5580  |
| S14 | 17.2      | 0.6406  |
| S15 | 26.4      | 0.1544  |
| S16 | 26.4      | 0.1528  |
| S17 | 19.4      | 0.4967  |
| S18 | 16.8      | 0.6671  |
| S19 | 21.8      | 0.3506  |
| S20 | 20.2      | 0.4430  |
| S21 | 23.0      | 0.2878  |
| S22 | 28.4      | 0.0993  |
| S23 | 17.0      | 0.6512  |
| S24 | 21.5      | 0.3703  |
| S25 | 23.4      | 0.2696  |
| S26 | 17.6      | 0.6120  |
| S27 | 16.7      | 0.6732  |
| S28 | 27.8      | 0.1151  |
| S29 | 16.3      | 0.6983  |
| S30 | 22.4      | 0.3201  |

Table 3: Mahalanobis distance (with unadjusted p-values) of each sample from the center of 20-dimensional principal component space.

```
[1] xtable_1.8-8          ClassDiscovery_3.4.10 oompaBase_3.2.11
[4] cluster_2.1.8.2
```

loaded via a namespace (and not attached):

```
[1] compiler_4.6.0 mclust_6.1.2    oompaData_3.1.5 tools_4.6.0
```