PCA, Mahalanobis Distance, and Outliers

Kevin R. Coombes

4 November 2011

Contents

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)
```

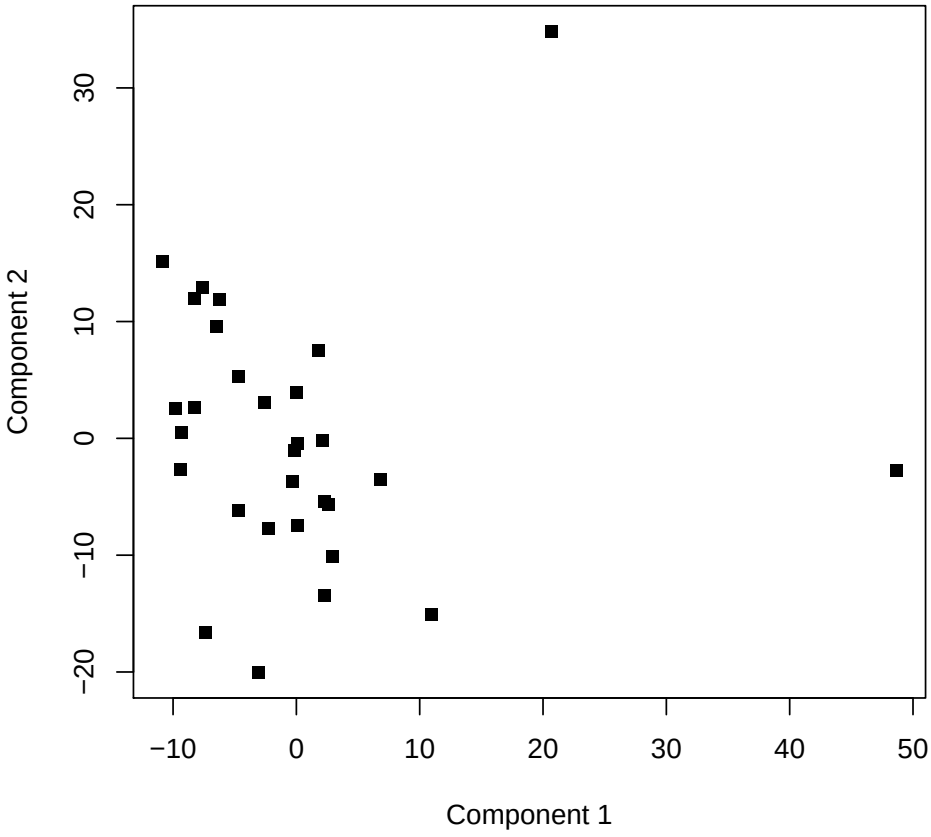

Figure 1: Principal components plot of the samples.

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 0.59
[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	49.7	0.0000	53.9	0.0000	54.9	0.0000	58.6	0.0000
S2	18.4	0.0001	25.6	0.0001	28.7	0.0014	31.3	0.0515
S3	0.0	0.9832	0.5	0.9914	9.5	0.4875	23.7	0.2558
S4	0.9	0.6396	1.9	0.8668	3.5	0.9659	20.4	0.4331
S5	0.5	0.7983	0.7	0.9802	1.7	0.9982	28.4	0.1008
S6	0.5	0.7717	3.9	0.5658	13.2	0.2111	20.1	0.4489
S7	2.8	0.2439	13.9	0.0162	18.3	0.0495	27.2	0.1287
S8	0.4	0.8225	3.6	0.6043	11.9	0.2914	20.0	0.4579
S9	0.8	0.6742	1.2	0.9492	8.9	0.5379	23.6	0.2604
S10	0.3	0.8721	7.9	0.1625	12.8	0.2354	24.1	0.2390
S11	0.6	0.7522	3.7	0.5951	16.3	0.0927	21.3	0.3772
S12	0.7	0.6945	5.3	0.3806	11.9	0.2901	20.4	0.4341
S13	0.4	0.8002	0.9	0.9677	8.8	0.5556	16.2	0.7041
S14	0.5	0.7910	2.6	0.7620	6.9	0.7354	18.3	0.5653
S15	0.0	0.9956	0.3	0.9984	10.5	0.4005	15.6	0.7382
S16	2.8	0.2497	5.3	0.3845	11.7	0.3033	23.9	0.2446
S17	1.5	0.4640	5.3	0.3784	7.7	0.6574	14.5	0.8041
S18	3.7	0.1610	6.0	0.3101	6.7	0.7548	17.8	0.6010
S19	1.5	0.4841	3.4	0.6431	13.7	0.1876	21.6	0.3622
S20	0.5	0.7862	2.5	0.7730	7.7	0.6592	20.7	0.4166
S21	1.7	0.4268	10.7	0.0577	16.5	0.0850	22.9	0.2925
S22	0.0	0.9992	3.7	0.5967	4.8	0.9032	11.5	0.9309
S23	0.1	0.9393	1.4	0.9191	6.0	0.8175	18.5	0.5554
S24	0.1	0.9406	3.6	0.6124	7.2	0.7049	23.1	0.2822
S25	0.7	0.7176	9.3	0.0966	14.2	0.1621	24.9	0.2040
S26	0.1	0.9468	1.7	0.8920	8.7	0.5580	18.9	0.5279
S27	1.8	0.4036	5.6	0.3464	9.3	0.4998	16.6	0.6782
S28	0.3	0.8589	2.4	0.7982	3.4	0.9713	23.6	0.2589
S29	1.1	0.5910	1.7	0.8940	7.6	0.6711	16.0	0.7162
S30	2.8	0.2424	2.9	0.7218	4.2	0.9364	21.2	0.3860

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 0.08 0.13 0.17 0.20 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)
```

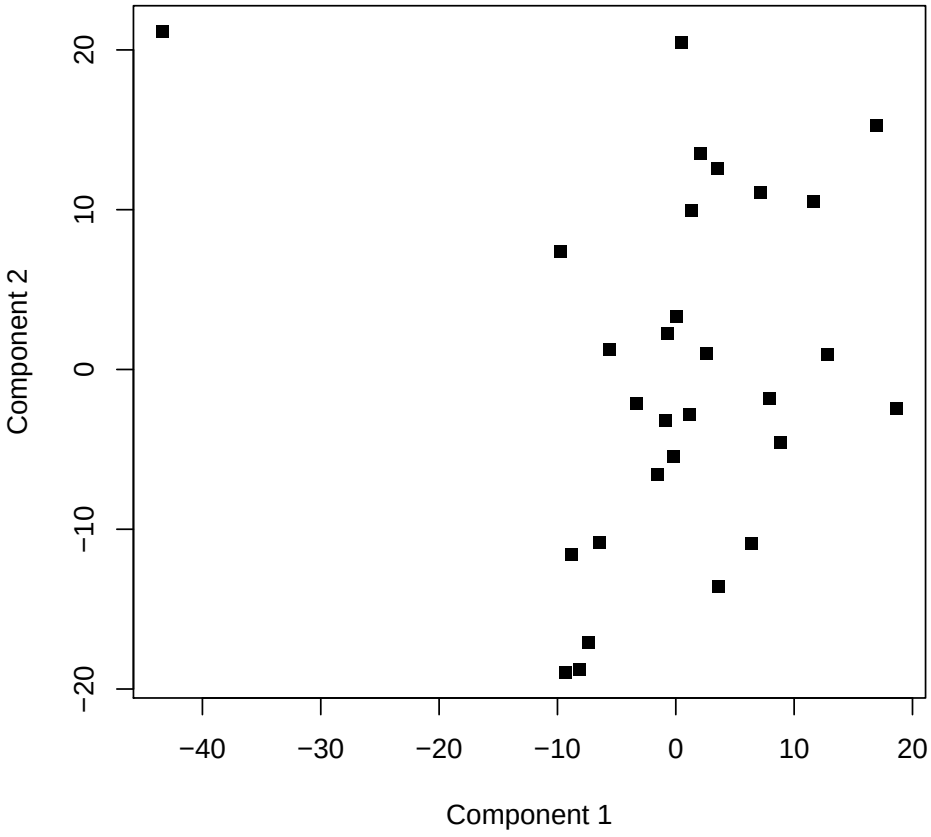

Figure 2: Principal components plot of the normal control samples, after omitting an extreme outlier.

	statistic	p.value
S2	42.2	0.0026
S3	23.9	0.2462
S4	21.5	0.3673
S5	27.7	0.1172
S6	20.4	0.4303
S7	25.1	0.1968
S8	19.7	0.4793
S9	25.0	0.2029
S10	23.9	0.2459
S11	20.4	0.4309
S12	19.4	0.4976
S13	15.3	0.7576
S14	17.4	0.6268
S15	28.0	0.1098
S16	23.9	0.2477
S17	18.5	0.5515
S18	17.3	0.6353
S19	22.3	0.3220
S20	20.1	0.4529
S21	23.2	0.2800
S22	11.8	0.9220
S23	17.8	0.5993
S24	22.3	0.3223
S25	24.0	0.2445
S26	18.2	0.5733
S27	16.6	0.6807
S28	23.3	0.2722
S29	15.7	0.7365
S30	23.1	0.2850

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/Kevin/AppData/Local/Temp/RtmpyiSwGj/Rbuild7e81be1a3c/ClassDiscovery/vignettes"
```

This analysis was performed in the following software environment:

```
> sessionInfo()

R version 3.5.1 (2018-07-02)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 7 x64 (build 7601) Service Pack 1

Matrix products: default

locale:
[1] LC_COLLATE=C                           LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252 LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

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

other attached packages:
[1] xtable_1.8-3         ClassDiscovery_3.3.9 oompaBase_3.2.6      cluster_2.0.7-1

loaded via a namespace (and not attached):
[1] compiler_3.5.1  mclust_5.4.1    oompaData_3.1.1 tools_3.5.1
```

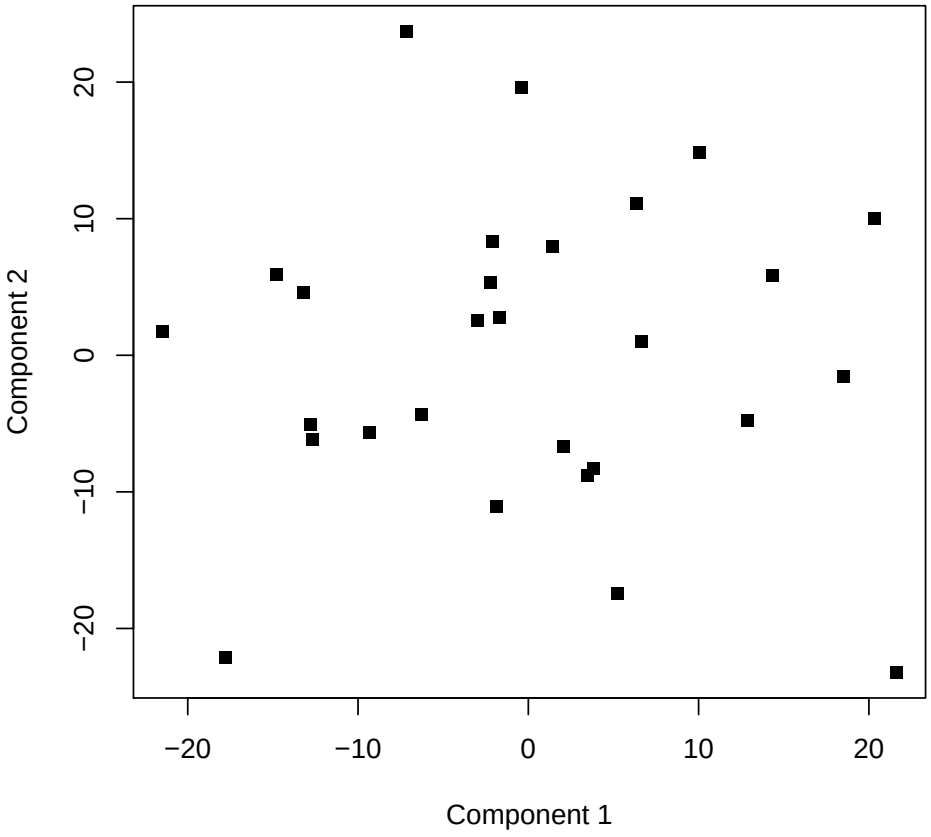

Figure 3: Principal components plot of the normal control samples, after omitting an extreme outlier.

	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.