Finding the Number of Principal Components

Min Wang[1], Steven M. Kornblau[2], and Kevin R. Coombes[3]

[1]Mathematical Biosciences Institute, The Ohio State University
[2]Dept. of Leukemia, The University of Texas MD Anderson Cancer Center
[3]Dept. of Biomedical Informatics, The Ohio State University

May 17, 2018

Contents

1 Getting Started

In this section, we describe how to select significant number of PCs using the **PCDimension** R package. The latest version of the package is always available from the R-Forge webpage (`http://r-forge.r-project.org/R/?group_id=1900`); the latest stable version can be found on CRAN. We illustrate the methods by exploring a small simulated data set.

First, we load all of the R library packages that we need for this analysis. Note that **PCDimension** implements the Broken-Stick model, the randomization-based procedure of ter Brack [1,2], and the Auer-Gervini model [3], while **nFactors** (developed by Raiche and Magis) is used to run Bartlett's test and its variants.

```
> library(PCDimension)
> library(nFactors) # implements Bartlett's test
> library(MASS) # for mvrnorm to simulate data
```

2 Testing on Unstructured Data

Next, we simulate an unstructured data set with random noise. That is, the variation of the data is isotropic and the number of significant PCs is 0. The data is generated via the command:

```
> set.seed(12345)
> NC <- 15
> NS <- 200
> ranData <- matrix(rnorm(NS*NC, 6), ncol=NC)
```

2.1 Bartlett's Test

Now, we apply Bartlett's test to the simulated data. The required input includes the raw data and the number of subjects (rows).

```
> nBartlett(data.frame(ranData), nrow(ranData))
```

```
bartlett anderson   lawley
      15       15        0
```

The original version of Bartlett's test, and the Anderson variant, fail to return the correct number of components. The Lawley variant does yield the correct value, 0.

2.2 Randomization-Based Methods

The **PCDimension** package implements both of the randomization-based statistics that were identified as successful in a previous study [4]. The number of permutations (default, $B = 1000$) and the significance level (default, $\alpha = 0.05$) are optional input arguments in addition to the required data set. The estimated number of PCs is the last point at which the p-value of the statistic of interest greater than the observed one is at least the threshold significance level.

```
> rndLambdaF(ranData) # input argument is data
```

```
rndLambda      rndF
        0         0
```

The randomization-based procedure successfully recovers the true number of PCs.

2.3 Broken-Stick

The **PCDimension** package also implements the Broken-Stick model. Both this model and the Auer-Gervini model require the eigenvalues from the singular value decomposition of the data matrix used to compute the principal components. We compute the decomposition using the `SamplePCA` function from the **ClassDiscovery** package, and then extract the variances.

```
> spca <- SamplePCA(t(ranData))
> lambda <- spca@variances[1:(NC-1)]
> bsDimension(lambda)
```

```
[1] 0
```

In the Broken-Stick model, the individual percentages of variance of the components are compared with the values expected from the "broken stick" distribution. The two distributions are compared element-by-element, and first value $d + 1$ where the expected value is larger than the observed value determines the dimension. The Broken-Stick model also correctly finds that there are zero significant PCs.

Note: In our implementation of tbhe `bsDimension` function, we add an extra parameter (`FUZZ`, with default value 0.005) for this comparison to deal with numerical errors in the estimates of the eigenvalues.

2.4 Auer-Gervini

We now use the `SamplePCA` object to construct an Auer-Gervini object.

```
> ag.obj <- AuerGervini(spca)
> agDimension(ag.obj)
```

```
[1] 0
```

The `agDimension` function takes an optional argument, `agfun` that specifies the method used to automate the computation of the number of PCs. The default value uses the **TwiceMean** method, which correctly concludes that there are zero significant PCs. We can also compare the results of multiple algorithms to automate the procedure.

```
> f <- makeAgCpmFun("Exponential")
> agfuns <- list(twice=agDimTwiceMean, specc=agDimSpectral,
+                km=agDimKmeans, km3=agDimKmeans3,
+                tt=agDimTtest, tt2=agDimTtest2,
+                cpt=agDimCPT, cpm=f)
> compareAgDimMethods(ag.obj, agfuns) # compare the list of all criteria
```

```
twice specc    km   km3    tt   tt2   cpt   cpm
    0     0     0     0     1     0     0     0
```

Overall, the Auer-Gervini model does an excellent job in selecting the actual number of components since 7 criteria out of 8 return 0 while only the "Ttest" procedure yields 1. If the majority rule is applied, that is, the estimated number of PCs is the one selected in more than 4 criteria in Auer-Gervini model, then we will definitely have 0 as the estimated number of components.

To get a more comprehensive understanding of the Broken-Stick method and the Auer-Gervini model, we use the command to generate the plots of these two models in Figures 1 and 2:

```
> bs <- brokenStick(1:NC, NC)
> bs0 <- brokenStick(1:(NC-1), (NC-1))
```

Figure 1 shows the broken stick distributions and the relative proportions of the variation that are explained by all the components in the simulated data set. The blue dotted line reprensents the broken stick distributions under the condition that p equals the number

```
> pts <- screeplot(spca, ylim=c(0, 0.2))
> lines(pts, bs, type='b', col='blue', lwd=2, pch=16)
> lines(pts[-NC], bs0, type='b', col='red', lwd=2, pch=16)
```

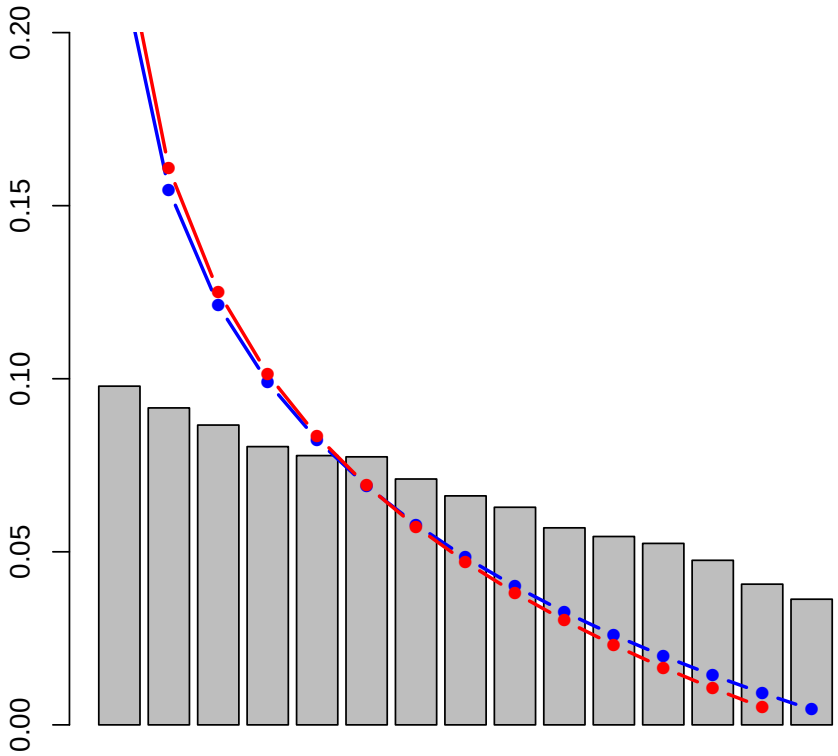

Figure 1: Screeplot of the data and the Broken-Stick model.

```
> plot(ag.obj, agfuns)
```

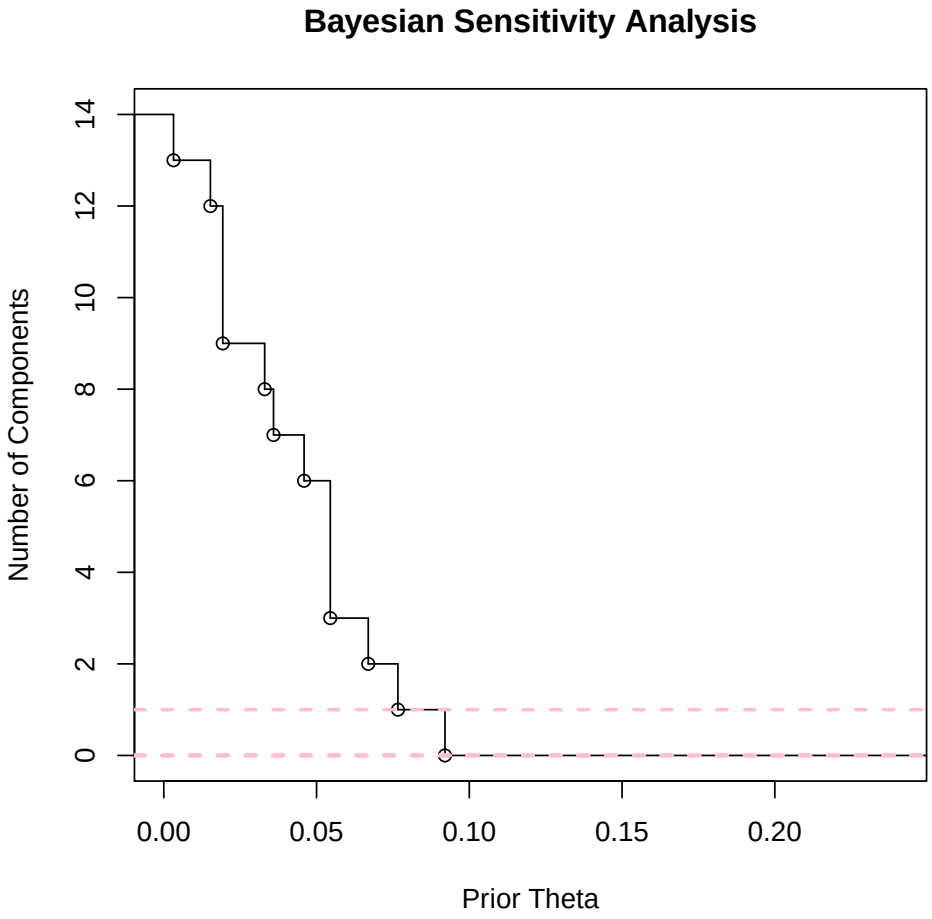

Figure 2: Auer-Gervini step function relating the prior hyperparameter θ to the maximum posterior estimate of the number K of significant principal components.

of features, while the red one means the broken stick distributions after removing the 0 eigenvalue with p equating with the number of features minus one. And there is almost no difference after removing the effect of eigenvalue 0. The grey rectangles are the relative proportions of the variation explained by all PCs. This figure provides a clear illustration on how the relative proportions are compared with the broken stick distributions and how the estimated number of PCs is chosen from the Broken-Stick model. Figure 2 illustrates how the Auer-Gervini model works. For the simulated data set, there are $NC = 15$ features, so the possible models $\mathcal{M}_d$ range from $\mathcal{M}_0$ to $\mathcal{M}_{14}$. The values $d = 0, 3, 4, 6, 7, 8, 11, 12, 13$ and 14 should be retained since the step functions are flat on those vertical coordinates. From the plot, we can see that the highest dimension d for which the step is significantly large is at $d = 0$. So $\mathcal{M}_0$ is a reasonable model.

3 Testing on Data with Structure

Now we simulate another dataset, with two groups of correlated samples.

```
> mu <- rep(0, 15)
> sigma <- matrix(0, 15, 15)
> sigma[1:8, 1:8] <- 0.7
> sigma[9:15, 9:15] <- 0.3
> diag(sigma) <- 1
> struct <- mvrnorm(200, mu, sigma)
```

3.1 Bartlett's Test

As before, we start by applying Bartlett's test to the simulated data.

```
> nBartlett(data.frame(struct), nrow(struct))
```

```
bartlett anderson   lawley
       7        7        8
```

All three variants grossly overestimate the dimension.

3.2 Randomization-Based Methods

Next, we apply ter Braak's randomization procedures. Here we again use the defaulkt values of the parametes, but include them explicitly to illustrate the function.

```
> rndLambdaF(struct, B = 1000, alpha = 0.05) # input argument is data

rndLambda      rndF
        2         9
```

The randomization-based procedure `rndLambda` successfully recovers the true number of PCs, but the `rndF` procedure overestimates it. These results are consistent with a larger set of simulations that we have performed..

3.3 Broken-Stick

Next, we apply the broken-stick model. As before, we first compute the singular value decomposition using the `SamplePCA` function from the **ClassDiscovery** package.

```
> spca <- SamplePCA(t(struct))
> lambda <- spca@variances[1:(NC-1)]
> bsDimension(lambda)

[1] 2
```

The scree-plot (Figure 3) explains how the broken-stick model gets the correct answer.

```
> bs <- brokenStick(1:NC, NC)
> bs0 <- brokenStick(1:(NC-1), (NC-1))
```

3.4 Auer-Gervini

We now use the `SamplePCA` object to construct an Auer-Gervini object.

```
> ag.obj <- AuerGervini(spca)
> agDimension(ag.obj)

[1] 2
```

The `agDimension` function takes an optional argument, `agfun` that specifies the method used to automate the computation of the number of PCs. The default value uses the **TwiceMean** method, which correctly concludes that there are zero significant PCs. We can also compare the results of multiple algorithms to automate the procedure.

```
> pts <- screeplot(spca)
> lines(pts, bs, type='b', col='blue', lwd=2, pch=16)
> lines(pts[-NC], bs0, type='b', col='red', lwd=2, pch=16)
```

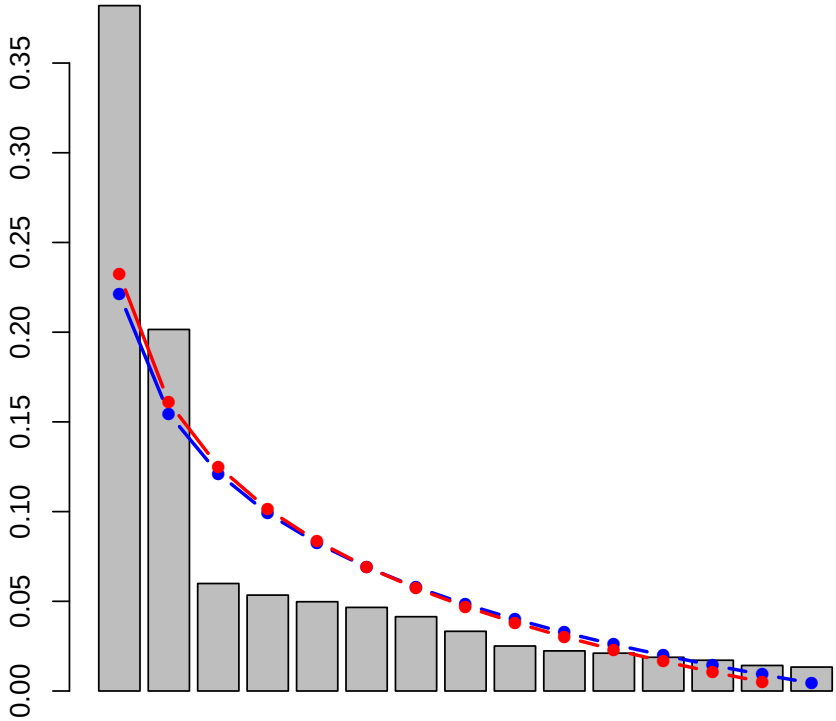

Figure 3: Screeplot of the structured data and the Broken-Stick model.

```
> f <- makeAgCpmFun("Exponential")
> agfuns <- list(twice=agDimTwiceMean, specc=agDimSpectral,
+                km=agDimKmeans, km3=agDimKmeans3,
+                tt=agDimTtest, tt2=agDimTtest2,
+                cpt=agDimCPT, cpm=f)
> compareAgDimMethods(ag.obj, agfuns) # compare the list of all criteria

twice specc    km   km3    tt   tt2   cpt   cpm
    2     2     2     2     8     2     2     8
```

Again, most of the criteria used to automate the Auer-Gervni method get te right answer. Two of them (`Ttest` and `CPM`), however, grossly overestimate it. Figure 4 shows the Auer-Gervini plot.

```
> plot(ag.obj, agfuns)
```

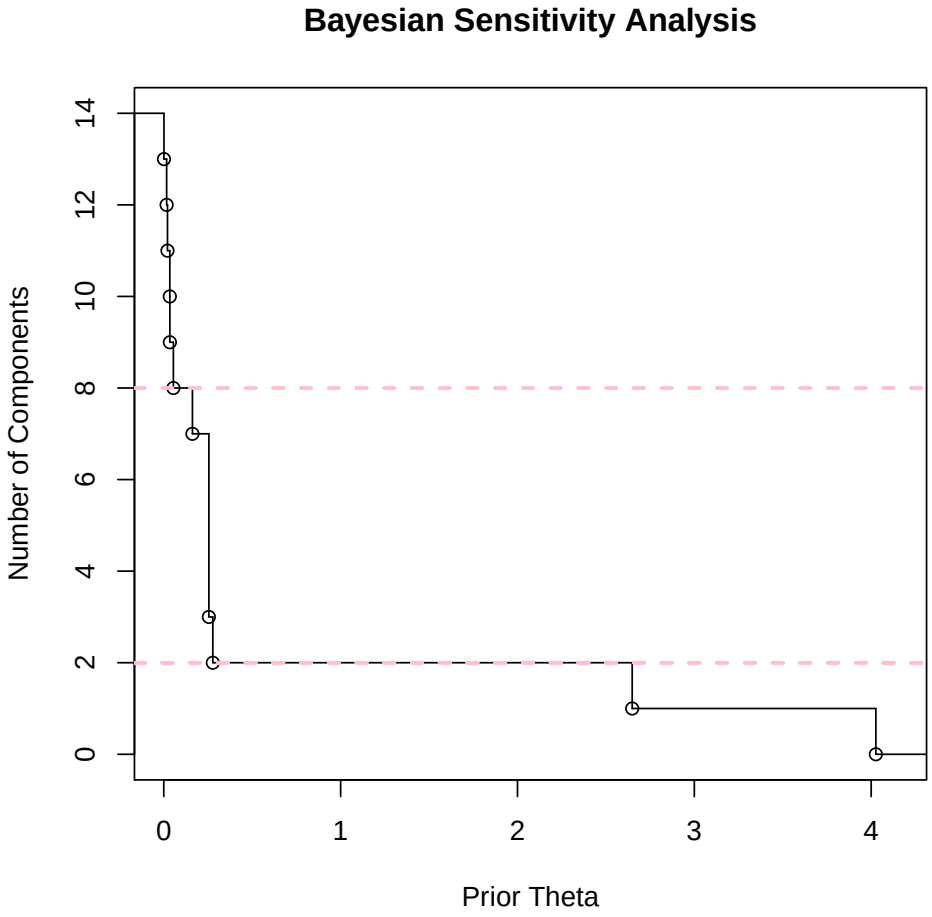

Figure 4: Auer-Gervini step function relating the prior hyperparameter θ to the maximum posterior estimate of the number K of significant principal components in the example with structured data.

4 References

[1] ter Braak CFJ. *CANOCO – a Fortran program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal component analysis and redundancy analysis (version 2.1).* Agricultural Mathematics Group, Report LWA-88-02, Wageningen, 1988.

[2] ter Braak CFJ. *Update notes: CANOCO (version 3.1).* Agricultural Mathematics Group, Wageningen, 1990.

[3] Auer P and Gervini D. Choosing principal components: A new graphical method based on Bayesian model selection. *Communications in Statistics - Simulation and Computation* 2008; 37: 962–977.

[4] Peres-Neto PR, Jackson DA and Somers KM. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis* 2005; 49: 974–997.