PCA test (version 2)

Dasapta Erwin Irawan

22/09/2014

*Note from 09/19 discussion:*

* This version throws out the "Q" variable.
* Use Random Forest package as comparison.

The PCA test was done using the pcaMethods package written by Stacklies et.al (2007). This package can automatically impute the missing values in the dataset.

The pcaMethods package provides several pca methods:

* svd: Singular value decomposition or classical pca similar to prcomp() or princomp(), for linear dataset
* svdimpute: svd with imputation (filling missing data)
* nipals: pca with iterative partial least squares method, for non-linear dataset
* bpca: pca using Bayesian model to handle NAs
* ppca: pca using probabilistic model to handle NAs

We applied the pca on the dataset, which have been divided in to three groups:

* Group 1: Physical parameters
* Group 2: Cations
* Group 3: Anions

For each group we used the following work flow:

* Run PCA using svdImpute.
* Retaining eigenvalues via function sDev() and plotting.
* Printing pca summary.
* Printing and plotting variable loadings to PCs.
* Printing and plotting cases scores to PCs.

# 0. Loading libraries and preparing dataset

* Installing pcaMethods package

source("http://bioconductor.org/biocLite.R")  
biocLite("pcaMethods")

* Loading libraries

require("pcaMethods") # PCA

## Loading required package: pcaMethods  
## Loading required package: Biobase  
## Loading required package: BiocGenerics  
## Loading required package: parallel  
##   
## Attaching package: 'BiocGenerics'  
##   
## The following objects are masked from 'package:parallel':  
##   
## clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,  
## clusterExport, clusterMap, parApply, parCapply, parLapply,  
## parLapplyLB, parRapply, parSapply, parSapplyLB  
##   
## The following object is masked from 'package:stats':  
##   
## xtabs  
##   
## The following objects are masked from 'package:base':  
##   
## Filter, Find, Map, Position, Reduce, anyDuplicated, append,  
## as.data.frame, as.vector, cbind, colnames, do.call,  
## duplicated, eval, evalq, get, intersect, is.unsorted, lapply,  
## mapply, match, mget, order, paste, pmax, pmax.int, pmin,  
## pmin.int, rank, rbind, rep.int, rownames, sapply, setdiff,  
## sort, table, tapply, union, unique, unlist  
##   
## Welcome to Bioconductor  
##   
## Vignettes contain introductory material; view with  
## 'browseVignettes()'. To cite Bioconductor, see  
## 'citation("Biobase")', and for packages 'citation("pkgname")'.  
##   
## Loading required package: Rcpp  
##   
## Attaching package: 'pcaMethods'  
##   
## The following object is masked from 'package:stats':  
##   
## loadings

require("pander") # output table formatting

## Loading required package: pander

require("dplyr") # table manipulation

## Loading required package: dplyr  
##   
## Attaching package: 'dplyr'  
##   
## The following objects are masked from 'package:BiocGenerics':  
##   
## intersect, setdiff, union  
##   
## The following objects are masked from 'package:stats':  
##   
## filter, lag  
##   
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

* Loading and grouping dataset (excluding Q)

data <- as.data.frame(read.csv("0806alldata.csv", header = TRUE))  
  
attach(data)

## The following object is masked from package:datasets:  
##   
## CO2

group1 <- data[,c("x", "y", "type",   
 "ec", "elv",   
 "ph", "hard",   
 "tds", "temp",  
 "eh", "cumrain",   
 "lag1")]  
  
group2 <- data[,c("x", "y", "type",   
 "ec", "elv",  
 "aq", "Ca", "Mg",   
 "Fe", "Mn", "K",   
 "Na", "cumrain",   
 "lag1")]  
  
group3 <- data[,c("x", "y", "type",   
 "ec", "elv",  
 "aq", "CO3",   
 "HCO3",   
 "CO2", "Cl",   
 "SO4", "NO2",   
 "NO3", "SiO2",   
 "cumrain",   
 "lag1")]

# 1. Run the PCA on Group1

* Run PCA using svdImpute

pcaSvd1 <- pca(group1,   
 method = "svdImpute",   
 scale = "uv",  
 center = T,  
 nPcs = 5,  
 evalPcs = 1:5)

According to the table, the first two PCs explain 87.5% of the variance. Major loadings on PC 1 are: y coordinate (0.1), Redox potential/eh (-0.97), and Hardness/hard (-0.14), while the ones on PC 2 are: TDS (0.49), elevation (-0.47), EC (0.45), and Temp (0.4).

Table 1 Summary and loadings of Group1

pander(summary(pcaSvd1))

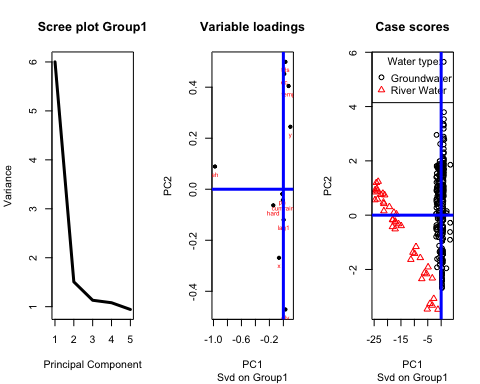
## svdImpute calculated PCA  
## Importance of component(s):  
## PC1 PC2 PC3 PC4 PC5  
## R2 0.827 0.04781 0.02663 0.02424 0.01845  
## Cumulative R2 0.827 0.87479 0.90142 0.92566 0.94412  
##   
## ---------------------------------------------------------  
## &nbsp; PC1 PC2 PC3 PC4 PC5   
## ------------------- ----- ------- ------- ------- -------  
## \*\*R2\*\* 0.827 0.04781 0.02663 0.02424 0.01845  
##   
## \*\*Cumulative R2\*\* 0.827 0.8748 0.9014 0.9257 0.9441   
## ---------------------------------------------------------

pander(loadings(pcaSvd1))

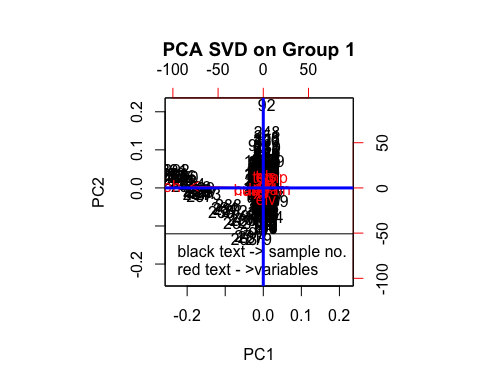
##   
## -----------------------------------------------------------  
## &nbsp; PC1 PC2 PC3 PC4 PC5   
## ------------- --------- -------- -------- -------- --------  
## \*\*x\*\* -0.06225 -0.2689 -0.1835 0.1639 -0.1426   
##   
## \*\*y\*\* 0.1001 0.2449 -0.04294 -0.1801 0.3813   
##   
## \*\*ec\*\* 0.01227 0.4519 -0.00272 0.1701 -0.5058   
##   
## \*\*elv\*\* 0.02926 -0.4713 0.07419 0.2948 -0.5244   
##   
## \*\*ph\*\* -0.01648 -0.01863 -0.7736 0.02017 -0.0967   
##   
## \*\*hard\*\* -0.144 -0.06319 -0.1382 0.01532 -0.01812  
##   
## \*\*tds\*\* 0.02951 0.4993 -0.08007 0.03277 -0.1452   
##   
## \*\*temp\*\* 0.0746 0.4046 -0.05642 0.08026 -0.2943   
##   
## \*\*eh\*\* -0.9786 0.08911 0.03193 -0.01143 0.005555  
##   
## \*\*cumrain\*\* -0.006632 -0.04233 0.3889 -0.6342 -0.3933   
##   
## \*\*lag1\*\* 0.003556 -0.1195 -0.4237 -0.6437 -0.1769   
## -----------------------------------------------------------

The following plots show the mixed plot of sample number distribution and variable in respect to the PCs. Here we can see that most of the groundwater samples fall along the PC 1, while the river water samples fall diagonally. The river water samples might have been shifted further from PC 1 towards PC 2. The shifting might be related to the high EH values which can induce oxydation processes in the river water. Most of the groundwater samples are located closer to PC1 in vertical orientation, similar to the distribution of most of the variables.

plot.new()  
par(mfrow=c(1,3))  
plot(sDev(pcaSvd1), type = 'l',   
 main = "Scree plot Group1",   
 xlab = "Principal Component",   
 ylab = "Variance",   
 lwd = 3)  
  
plot(loadings(pcaSvd1),   
 pch = 20,  
 main = "Variable loadings",  
 sub = "Svd on Group1")  
text(loadings(pcaSvd1),   
 row.names(loadings(pcaSvd1)),  
 cex=0.6,   
 pos=1,   
 col="red")  
abline(h = 0, v = 0, col = "blue", lwd = 3)  
  
plot(scores(pcaSvd1),   
 pch = c(group1$type),   
 col = c(group1$type),  
 main = "Case scores",  
 sub = "Svd on Group1")  
legend("topleft",   
 c("Groundwater", "River Water"),  
 title = "Water type:",  
 pch = c(1, 2),   
 col = c("black", "red"))  
abline(h = 0, v = 0, col = "blue", lwd = 3)



plot.new()  
par(mfrow=c(1,1))  
biplot(pcaSvd1,   
 main = "PCA SVD on Group 1")  
legend("bottomright", c("black text -> sample no.", "red text - >variables"))  
abline(h = 0, v = 0, col = "blue", lwd = 3)

 Figure 1 Plots of Group1: Scree, loadings, scores, and biplot

# 2. Run the PCA on Group2

pcaSvd2 <- pca(group2,   
 method = "svdImpute",   
 scale = "uv",  
 center = T,  
 nPcs = 5,  
 evalPcs = 1:5)

In the following table, we can see the summary, loadings, and scores table for Group 2 PCA. The first two PCs explains only 42% of the total variance. The major loadings for PC1 are: Ca (0.42), Mg (0.42), and Na (0.41), while for PC2 are: x coordinate (0.44), y coordinate (-0.41), Mn (-0.58).

Table 2 Summary and loadings of Group2

pander(summary(pcaSvd2))

## svdImpute calculated PCA  
## Importance of component(s):  
## PC1 PC2 PC3 PC4 PC5  
## R2 0.2812 0.1402 0.09752 0.08543 0.07816  
## Cumulative R2 0.2812 0.4214 0.51889 0.60432 0.68249  
##   
## ---------------------------------------------------------  
## &nbsp; PC1 PC2 PC3 PC4 PC5   
## ------------------- ------ ------ ------- ------- -------  
## \*\*R2\*\* 0.2812 0.1402 0.09752 0.08543 0.07816  
##   
## \*\*Cumulative R2\*\* 0.2812 0.4214 0.5189 0.6043 0.6825   
## ---------------------------------------------------------

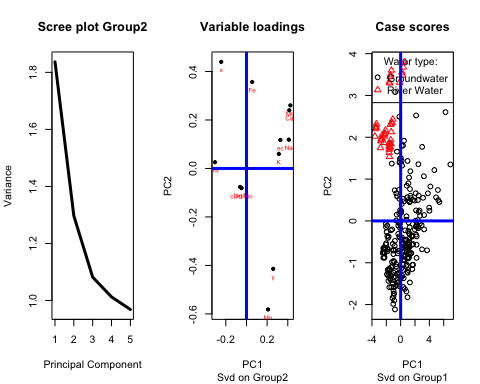
pander(loadings(pcaSvd2))

##   
## ----------------------------------------------------------  
## &nbsp; PC1 PC2 PC3 PC4 PC5   
## ------------- -------- -------- -------- -------- --------  
## \*\*x\*\* -0.2448 0.4397 -0.06425 0.0613 -0.1993   
##   
## \*\*y\*\* 0.2585 -0.4141 -0.03813 -0.1193 -0.1569   
##   
## \*\*ec\*\* 0.3283 0.1173 0.1852 -0.1147 0.4473   
##   
## \*\*elv\*\* -0.305 0.02586 0.2125 -0.04192 0.03601   
##   
## \*\*Ca\*\* 0.4151 0.2397 -0.1301 0.07614 -0.1493   
##   
## \*\*Mg\*\* 0.4252 0.2601 0.02015 0.0607 0.1441   
##   
## \*\*Fe\*\* 0.05489 0.3568 0.0935 -0.7376 0.1646   
##   
## \*\*Mn\*\* 0.2092 -0.5817 0.1862 -0.2229 0.04176   
##   
## \*\*K\*\* 0.3137 0.05992 -0.2743 0.1778 -0.4328   
##   
## \*\*Na\*\* 0.41 0.1189 0.04121 0.05847 -0.01283  
##   
## \*\*cumrain\*\* -0.04763 -0.08078 -0.5937 0.2422 0.6637   
##   
## \*\*lag1\*\* -0.06194 -0.07626 -0.6518 -0.5199 -0.1858   
## ----------------------------------------------------------

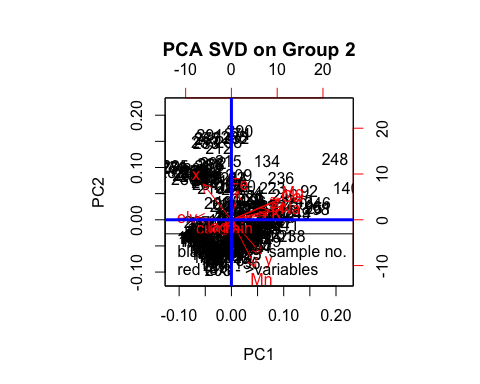
The following figures show the scree plot, loading, and scores plot of Group 2 PCA. Based on the first two PCs, we can see that most of the variables are distributed in the top-right quadrant. The rain-related variables, the cummulative rain and lag-1 rain, are located in the bottom-left quadrant. The spatial variables, x coordinate and elevation are located in the topleft quadrant. The y coordinate, however, is located in the same bottom-right quadrant as Manganese.

The groundwater samples are scattered in the top-right, bottom-righ and bottom-left corners of the plot. The river water samples are clustered in the top-left corner.

plot.new()  
par(mfrow=c(1,3))  
plot(sDev(pcaSvd2), type = 'l',   
 main = "Scree plot Group2",   
 xlab = "Principal Component",   
 ylab = "Variance",   
 lwd = 3)  
  
plot(loadings(pcaSvd2),   
 pch = 20,  
 main = "Variable loadings",  
 sub = "Svd on Group2")  
text(loadings(pcaSvd2),   
 row.names(loadings(pcaSvd2)),  
 cex=0.6,   
 pos=1,   
 col="red")  
abline(h = 0, v = 0, col = "blue", lwd = 3)  
  
plot(scores(pcaSvd2),   
 pch = c(group2$type),   
 col = c(group2$type),  
 main = "Case scores",  
 sub = "Svd on Group1")  
legend("topright",   
 c("Groundwater", "River Water"),  
 title = "Water type:",  
 pch = c(1, 2),   
 col = c("black", "red"))  
abline(h = 0, v = 0, col = "blue", lwd = 3)



plot.new()  
par(mfrow=c(1,1))  
biplot(pcaSvd2,   
 main = "PCA SVD on Group 2")  
legend("bottomright", c("black text -> sample no.", "red text - >variables"))  
abline(h = 0, v = 0, col = "blue", lwd = 3)

 Figure 2 Plots of Group2: Scree, loadings,scores, and biplot of Group2

# 3. Run the PCA on Group3

pcaSvd3 <- pca(group3,   
 method = "svdImpute",   
 scale = "uv",  
 center = T,  
 nPcs = 5,  
 evalPcs = 1:5)

In the following table, we can see the summary, loadings, and scores table for Group 2 PCA. The first two PCs explains only 33% of the total variance. The major loadings for PC1 are: Cl (0.49), HCO3 (0.43), EC (0.38), and elevation (-0.36), while for PC2 are: NO2 (0.58), NO3 (0.41), CO3 (-0.41), and y coordinate (0.39).

Table 3 Summary and loadings of Group3

pander(summary(pcaSvd3))

## svdImpute calculated PCA  
## Importance of component(s):  
## PC1 PC2 PC3 PC4 PC5  
## R2 0.1985 0.1331 0.09616 0.0824 0.07553  
## Cumulative R2 0.1985 0.3316 0.42777 0.5102 0.58570  
##   
## --------------------------------------------------------  
## &nbsp; PC1 PC2 PC3 PC4 PC5   
## ------------------- ------ ------ ------- ------ -------  
## \*\*R2\*\* 0.1985 0.1331 0.09616 0.0824 0.07553  
##   
## \*\*Cumulative R2\*\* 0.1985 0.3316 0.4278 0.5102 0.5857   
## --------------------------------------------------------

pander(loadings(pcaSvd3))

##   
## ------------------------------------------------------------  
## &nbsp; PC1 PC2 PC3 PC4 PC5   
## ------------- ---------- -------- -------- -------- --------  
## \*\*x\*\* -0.2742 0.2282 -0.1302 0.03249 0.3724   
##   
## \*\*y\*\* 0.3105 -0.3873 -0.2604 -0.2645 -0.01471  
##   
## \*\*ec\*\* 0.3799 0.1128 0.05207 0.3422 -0.02558  
##   
## \*\*elv\*\* -0.3609 -0.1355 -0.02988 0.3982 0.1473   
##   
## \*\*CO3\*\* -5.996e-05 -0.4104 0.09646 0.2608 0.1112   
##   
## \*\*HCO3\*\* 0.4288 0.163 -0.08358 -0.01585 -0.08037  
##   
## \*\*CO2\*\* 0.07661 -0.08233 0.5712 -0.06414 -0.3889   
##   
## \*\*Cl\*\* 0.4922 0.04829 0.04742 0.1927 0.1947   
##   
## \*\*SO4\*\* 0.3229 0.03797 0.2159 0.1208 0.413   
##   
## \*\*NO2\*\* -0.04244 0.5846 0.1889 0.1173 0.03898   
##   
## \*\*NO3\*\* -0.01683 0.4091 0.1256 -0.2588 -0.02617  
##   
## \*\*SiO2\*\* 0.1026 0.1019 -0.3589 -0.4817 0.231   
##   
## \*\*cumrain\*\* -0.01957 -0.1425 0.3096 -0.114 0.6308   
##   
## \*\*lag1\*\* -0.0675 -0.1467 0.4941 -0.451 0.08141   
## ------------------------------------------------------------

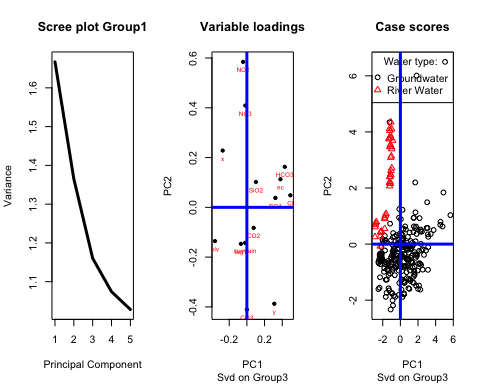
The following figures show the scree plot, loading, and scores plot of Group 2 PCA. Based on the first two PCs, we can see that most of the variables are distributed in the top-right quadrant. The HCO3, SiO2, SO4, and Cl are located in this quadrant, as well as EC.

The rain-related variables, the cummulative rain and lag-1 rain, are located in the bottom-left quadrant along with CO3, while the CO2 is shifted to bottom-right quadrant.

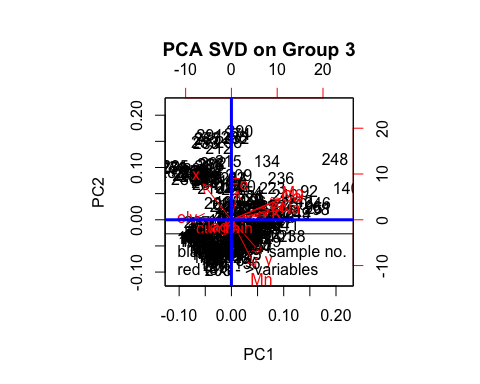
Both of NO2 and NO3 are located in the top-left quadrant. The spatial variables, x coordinate and elevation are located in the topleft quadrant. The spatial variables, x-y and elevation, are all located in the different quadrant.

Similar to Group 2, the groundwater samples of Group 3 are distributed in the top-right, bottom-right, and bottom-left corners, with high concentration in the centre of the plot. The river water are also located in the top-left corner in vertical orientation.

plot.new()  
par(mfrow=c(1,3))  
plot(sDev(pcaSvd3), type = 'l',   
 main = "Scree plot Group1",   
 xlab = "Principal Component",   
 ylab = "Variance",   
 lwd = 3)  
  
plot(loadings(pcaSvd3),   
 pch = 20,  
 main = "Variable loadings",  
 sub = "Svd on Group3")  
text(loadings(pcaSvd3),   
 row.names(loadings(pcaSvd3)),  
 cex=0.6,   
 pos=1,   
 col="red")  
abline(h = 0, v = 0, col = "blue", lwd = 3)  
  
plot(scores(pcaSvd3),   
 pch = c(group3$type),   
 col = c(group3$type),  
 main = "Case scores",  
 sub = "Svd on Group3")  
legend("topright",   
 c("Groundwater", "River Water"),  
 title = "Water type:",  
 pch = c(1, 2),   
 col = c("black", "red"))  
abline(h = 0, v = 0, col = "blue", lwd = 3)



plot.new()  
par(mfrow=c(1,1))  
biplot(pcaSvd2,   
 main = "PCA SVD on Group 3")  
legend("bottomright", c("black text -> sample no.", "red text - >variables"))  
abline(h = 0, v = 0, col = "blue", lwd = 3)

 Figure 3 Plots of Group3: Scree, loadings,scores, and biplot of Group3

# 4. Preliminary results

We can see the differentiation of groundwater samples from river water samples in terms of physical properties, cations and anions concentration.

The role of spatial variables (x-y and elevation) to PC1 and PC2, is not equal as shown by the loading values. The NO2 and NO3 concentrations as contaminant signature are mostly correlated with river water, while groundwater has more control from surficial process (CO2, CO3, HCO3), and from the geology (Fe, Mg, SiO2). The Cl and SO4 variables imprinted in the groundwater are the indications seepage from deeper aquifer.

# References

* [Introduction to pcaMethods package - Documentation](http://www.bioconductor.org/packages/release/bioc/vignettes/pcaMethods/inst/doc/pcaMethods.pdf)
* [Introduction to pcaMethods package - R script](http://www.bioconductor.org/packages/release/bioc/vignettes/pcaMethods/inst/doc/pcaMethods.R)
* R packages

citation("base")

##   
## To cite R in publications use:  
##   
## R Core Team (2014). R: A language and environment for  
## statistical computing. R Foundation for Statistical Computing,  
## Vienna, Austria. URL http://www.R-project.org/.  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Manual{,  
## title = {R: A Language and Environment for Statistical Computing},  
## author = {{R Core Team}},  
## organization = {R Foundation for Statistical Computing},  
## address = {Vienna, Austria},  
## year = {2014},  
## url = {http://www.R-project.org/},  
## }  
##   
## We have invested a lot of time and effort in creating R, please  
## cite it when using it for data analysis. See also  
## 'citation("pkgname")' for citing R packages.

citation("knitr")

##   
## To cite the 'knitr' package in publications use:  
##   
## Yihui Xie (2014). knitr: A general-purpose package for dynamic  
## report generation in R. R package version 1.6.  
##   
## Yihui Xie (2013) Dynamic Documents with R and knitr. Chapman and  
## Hall/CRC. ISBN 978-1482203530  
##   
## Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible  
## Research in R. In Victoria Stodden, Friedrich Leisch and Roger  
## D. Peng, editors, Implementing Reproducible Computational  
## Research. Chapman and Hall/CRC. ISBN 978-1466561595

citation("pcaMethods")

##   
## The pcaMethods package implement algorithms found in several  
## different publication. Refer to function documentation for  
## reference to the original articles.  
##   
## Stacklies, W., Redestig, H., Scholz, M., Walther, D. and Selbig,  
## J. pcaMethods -- a Bioconductor package providing PCA methods  
## for incomplete data. Bioinformatics, 2007, 23, 1164-1167  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Article{,  
## title = {pcaMethods -- a Bioconductor package providing PCA methods for incomplete data},  
## author = {Wolfram Stacklies and Henning Redestig and Matthias Scholz and Dirk Walther and Joachim Selbig},  
## journal = {Bioinformatics},  
## year = {2007},  
## pages = {1164--1167},  
## volume = {23},  
## }  
##   
## This free open-source software implements academic research by the  
## authors and co-workers. If you use it, please support the project  
## by citing the appropriate journal articles.