Handling of data containing outliers

Wolfram Stacklies and Henning Redestig
CAS-MPG Partner Institute for Computational Biology (PICB)
Shanghai, P.R. China
and
Max Planck Institute for Molecular Plant Physiology
Potsdam, Germany
http://bioinformatics.mpimp-golm.mpg.de/

May 2, 2019

1 PCA robust to outliers

Away from often showing missing values, Microarray or Metabolite data are often corrupted with extreme values (outliers). Standard SVD is highly susceptible to outliers. In the extreme case, an individual data point, if sufficiently outlying, can draw even the leading principal component toward itself. This problem can be addressed by using a robust analysis method. Hereto we provide `robustSvd`, a singular value decomposition robust to outliers. `robustPca` is a PCA implementation that resembles the original `R` `prcomp` method, with the difference that it uses `robustSvd` instead of the standard `svd` function.
Robust SVD and its application to microarray data were proposed in [1] and [2]. The algorithm is based on the idea to use a sequential estimation of the eigenvalues and left and right eigenvectors that ignores missing values and is resistant to outliers.
The `robustSvd` script included here was contributed by Kevin Wright. Thanks a lot to him!

2 Outliers and missing value imputation

The problem of outliers is similar to the missing data problem in the sense that extreme values provide no or wrong information. They are generally artifacts of the experiment and provide no information about the underlying biological processes.
Most of the PCA methods coming with the package were not designed to be robust to outliers in the sense that they will converge to the standard PCA solution on a complete data set. Yet, an applicable solution is to remove obvious outliers from the data first (by setting them NA) and to then estimate the PCA solution on the incomplete data. This is likely to produce accurate results if the number of missing data does not exceed a certain amount, less than 10% should be a good number.

The following example illustrates the effect of outliers and the use of robust methods. First, we attach the complete metabolite data set and create 5% outliers. We mean center the data before we create outliers because these large artificial outliers will strongly shift the original means. This would not allow for objective comparison between the differnt results obtained, e.g. when doing scatterplots.

```
> library(pcaMethods)

> data(metaboliteDataComplete)
> mdc          <- scale(metaboliteDataComplete, center=TRUE, scale=FALSE)
> cond         <- runif(length(mdc)) < 0.05
> mdcOut       <- mdc
> mdcOut[cond] <- 10
```

Then we calculate a PCA solution using standard SVD and robust SVD.

```
> resSvd       <- pca(mdc, method="svd", nPcs=5, center=FALSE)
> resSvdOut    <- pca(mdcOut, method="svd", nPcs=5, center=FALSE)
> resRobSvd    <- pca(mdcOut, method="robustPca", nPcs=5, center=FALSE)
```

Now we use `PPCA` to estimate the PCA solution, but set the outliers NA before.

```
> mdcNa        <- mdc
> mdcNa[cond]  <- NA
> resPPCA      <- pca(mdcNa, method="ppca", nPcs=5, center=FALSE)
```

To check the robustness to outliers we can just do a scatterplot comparing the results to the optimal PCA solution for the complete data set (which is `resSvd`). In Figure 1 we plot the estimated and original loadings against each other.

References

[1] Hawkins, D.M., Liu, L. and Young, S.S. *Robust Singular Value Decomposition.* National Institute of Statistical Sciences, 2001, Tech Report 122.

[2] Liu, L., Hawkins, D.M., Ghosh, S. and Young, S.S. *Robust singular value decomposition analysis of microarray data.* PNAS, 2003;100:13167–13172.

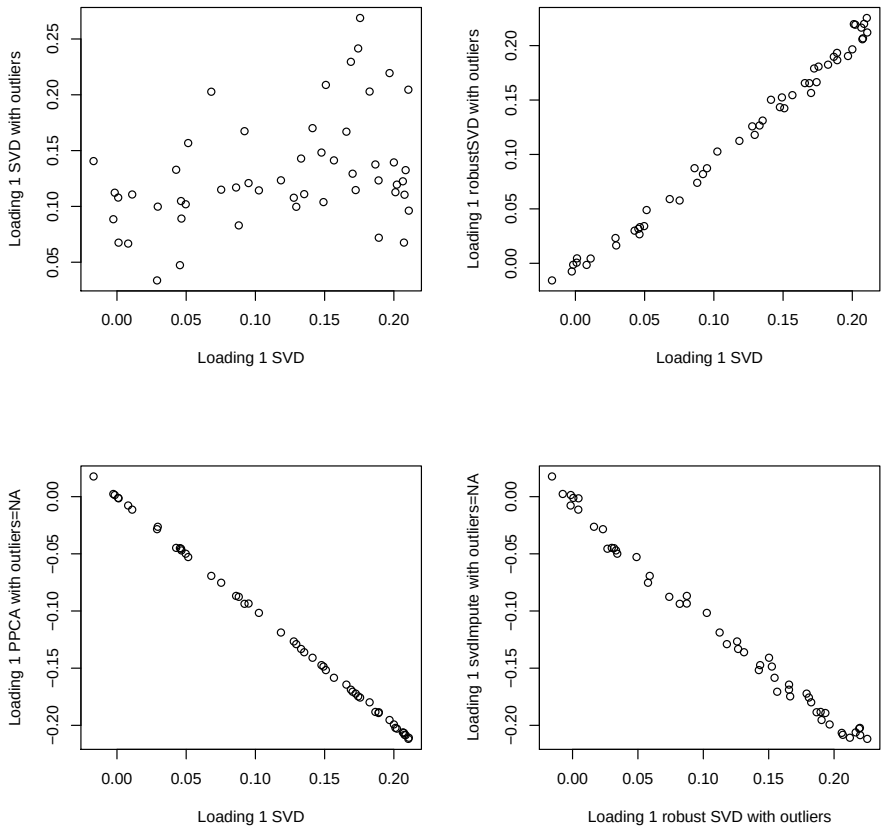

Figure 1: Figures show (from left to right):
Original PCA solution vs. solution on data with outliers;
Original PCA solution vs. robust PCA solution on data with outliers;
Original PCA solution vs. PPCA solution on data where outliers=NA;
Robust PCA solution vs. PPCA solution on data with outliers / outliers=NA.