R Package RobLoxBioC: Infinitesimally robust estimators for preprocessing gene expression data

Matthias Kohl[1]

1. Department of Mathematics, University of Bayreuth. Contact: Matthias.Kohl@uni-bayreuth.de

Keywords: gene expression, preprocessing, infinitesimal robustness, radius-minimax estimator

The preprocessing of gene expression data for several platforms routinely includes the aggregation of multiple raw signal intensities to a single expression value. Examples are the computation of a single expression measure based on the perfect match (PM) and miss match (MM) probes in case of the Affymetrix technology, the summarization of bead level values to bead summary values in case of the Illumina technology, or the aggregation of replicated measurements in case of other technologies including real-time quantitative polymerase chain reaction (RT-qPCR) platforms.
Our new package `RobLoxBioC` provides a way to use infinitesimally robust estimators (cf. Rieder (1994), Kohl (2005)) for this purpose. More precisely, we assume normal location and scale and envelop this (ideal) model with an infinitesimally (i.e., shrinking) contamination neighorhood (Tukey's gross error model) where the exact size/radius of the neighorhood is unknown. The optimally robust radius-minimax (rmx) estimators for this setup, minimizing the relative asymptotic minimax MSE for some given radius interval, can be read off from Rieder et al. (2008) and are implemented in our new package `RobLoxBioC`.
In case of Affymetrix data we implemented an algorithm which is similar to MAS 5.0 (cf. Affymetrix, Inc. (2002)). The main difference is the substitution of the Tukey one-step estimator by an rmx k-step ($k \geq 1$) estimator. The rmx estimators can also be applied to Illumina bead level data as well as to data from other platforms or other omics disciplines (e.g., Proteomics or Metabolomics) incorporating replicated measurements.
We will give some comparisons between the results obtained for our rmx estimators and estimators implemented in Bioconductor (cf. Gentleman et al. (2004)) using datasets from literature.

References

Affymetrix, Inc. (2002). *S*tatistical Algorithms Description Document. Affymetrix, Santa Clara.

R. Gentleman, V. J. Carey, D. M. Bates, B.Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *G*enome Biology, 5:R80.

M. Kohl (2005). *N*umerical Contributions to the Asymptotic Theory of Robustness. PhD-thesis, University of Bayreuth, Bayreuth, `http://www.stamats.de/ThesisMKohl.pdf`.

H. Rieder (1994). *R*obust Asymptotic Statistics. Springer, New York.

H. Rieder, M. Kohl and P. Ruckdeschel (2008). The Costs of not Knowing the Radius. *S*tat. Meth. & Appl., 17(1): 13–40.