# Choosing the Best Method for Gene Expression log- log Linear Models Using Multiple CART Trees

Martín Castillo, Rodrigo Assar
Institute of Biomedical Sciences, School of Medicine, University of Chile
mcastillo@dim.uchile.cl, rodrigo.assar@gmail.com

*Keywords: Log-log linear model estimations, Microarrays, RNA-Seqs.*

Microarray and RNA-Seq techniques are used to infer genes showing differential expressions on treatment conditions through the analysis of log-log linear models for the expression with treatment compared with control condition. Due to costs and technical limitations usually the experiments present small-sized samples and high contamination; therefore, choosing the estimation method for coefficients of such models becomes a challenge [1]. Herein, we simulate microarray and RNA-Seq experiments and analyze a log-log linear model with contaminations at both conditions, varying key features: the sample size $n$, contamination type (*light-tailed* or *heavy-tailed*), contamination proportion $p$, and error variance $\sigma^2$. For each features configuration we computed the accuracy at each method among least absolute deviations ($l_1$), ordinary least squares ($l_2$), and Huber M-Estimators ($HM$). Using this information, we built a machine learning that, based on classification CART trees [2], automatically decides the best method depending on simple questions. Restricted to light tails, $l_2$ leads if $n$ is small and $\sigma^2$ low, while $l_1$ leads if $\sigma^2$ is moderate and $p$ high, and for $p$ low and $\sigma^2$ moderate $HM$ leads. In case of heavy tails, $HM$ leads if $p$ is moderate. Simulation results on method decisions agree with theoretical analysis [3], adding more information for non-extreme conditions, and show good sensitivity and specificity in true experiments.

## References

[1] L. Cheng *et al.*, *Challenges and Strategies for Differential Transcriptome Analysis from Microarray to Deep Sequencing in Statistics*, Ann Biom Biostat **2(1)**, 2015.

[2] C. Hodar *et al.*, *Genome-wide identification of new Wnt/$\beta$-catenin target genes in the human genome using CART method*, BMC Genomics **11(1)** 348, 2010.

[3] S. Flores, *Sharp non-asymptotic performance bounds for $l_1$ and Huber robust regression estimators*, TEST 1–17, 2015.