Gerard, David; Stephens, Matthew

Summary: Unwanted variation, including hidden confounding, is a well-known problem in many fields, but particularly in large-scale gene expression studies. Recent proposals to use control genes, genes assumed to be unassociated with the covariates of interest, have led to new methods to deal with this problem. Several versions of these removing unwanted variation (RUV) methods have been proposed, including RUV1, RUV2, RUV4, RUVinv, RUVrinv, and RUVfun. Here, we introduce a general framework, RUV*, that both unites and generalizes these approaches. This unifying framework helps clarify the connections between existing methods. In particular, we provide conditions under which RUV2 and RUV4 are equivalent. The RUV* framework preserves an advantage of the RUV approaches, namely, their modularity, which facilitates the development of novel methods based on existing matrix imputation algorithms. We illustrate this by implementing RUVB, a version of RUV* based on Bayesian factor analysis. In realistic simulations based on real data, we found RUVB to be competitive with existing methods in terms of both power and calibration. However, providing a consistently reliable calibration among the data sets remains challenging.

MSC:
62P10 Applications of statistics to biology and medical sciences; meta analysis
92D20 Protein sequences, DNA sequences

Keywords:
batch effect; correlated test; gene expression; hidden confounding; negative control; RNA-seq; unobserved confounding; unwanted variation

Software:
leapp; missForest; Eigenstrat; FAMT; ruv; denoiseR

Full Text: DOI

References:


data by modeling hidden covariates with prior knowledge. PloS one 8, 1-10.


[63] Matthew Stephens Departments of Human Genetics and Statistics, University of Chicago, Chicago, IL 60637, USA. E-mail: mstephens@uchicago.edu (Received August 2018; accepted September 2019)

This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.