Abstract

I propose a novel variation (Pro-SVA) on iteratively reweighted surrogate variable analysis (IRW-SVA) for detecting and measuring batch effects in high dimensional gene expression data. Specifically, I propose to use the matrix-free high dimensional factor analysis (HDFA) algorithm instead of singular value decomposition (SVD) in the IRW-SVA iterations. HDFA efficiently provides the maximum likelihood estimates of the error variances and batch loadings, which can subsequently be used to estimate the batch factors. To evaluate the performance of Pro-SVA, I simulated 100 samples of 1,000 genes with batch effects and (1) no biological effects, (2) biological effects for half of the genes, or (3) biological effects for all genes. To compare the methods, I estimated the batch-induced correlation matrix using both methods and computed the relative Frobenius distance of this estimate to the true correlation matrix. The results show that Pro-SVA obtains better estimates of the correlation matrix than IRW-SVA in most cases, especially when there are no biological effects or when the biological covariate affects only half the genes. Therefore, Pro-SVA holds promise as a new approach to detect and account for batch effects in high-dimensional gene expression datasets.