Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2016:7**
http://www.math.su.se

Statistical process control for next-generation sequencing quality control data

Erik Thorsén*

December 2016

Abstract

Assuring quality output is a vital part of the management of production processes. This is no different to the area of sequencing where the quality of sequenced data needs to be assured. In this thesis we have provided a solution to the issue of detecting and finding changes in transformed next-generation sequencing (NGS) quality control data using control charts from statistical process control (SPC) together with a change-point estimation procedure. The transformed data was assumed to follow a multivariate normal distribution. We monitored the mean vector using Hotelling's T^2 statistic and Croiser's MCUSUM control chart (cf. Hotelling, (1947), Croiser, (1988)). To monitor the covariance matrix we made use of properties of the singular Wishart distribution, introduced by Bodnar et al., (2009). Change-points were estimated using a generalized likelihood ratio when Croiser's MCUSUM chart gave a indication of a change. Our model was applied to data from an individual machine with a certain setting. A simulation study was performed to test how the control charts performed under tran- sient and persistent changes. The results of this simulation showed that Hotelling's T^2 control chart was efficient for detecting transient and large changes but poorly for persistent and small. The MCUSUM control charts detected small and persistent changes in mean and covariance matrix while worse than Hotelling's T^2 for transient and large. In the simulation study, the change point estimation procedure was shown to be accurate for large persistent changes. The constructed control charts were also applied to transformed quality control data from other machines of the same sort. In this application, Hotelling's T^2 control chart showed no great difference for transformed quality control data between the machines but a difference between the settings on these machines. However, the MCUSUM charts detected large structural differences between the machines on the same setting. These differences were discovered in the mean and covariance matrix. All control charts and simulations where migrated to C++ using the Rcpp package together with OpenMP, a parallel programming model, to increase R's computational power. The improvement was shown to be large compared to base R performance in a benchmark.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: ethorsn@gmail.com. Supervisor: Taras Bodnar.