ロジスティック回帰における
共変量測定誤差をもつデータの統合

松田 孟留
東京大学大学院情報理工学系研究科
理化学研究所脳神経科学研究センター

Logistic regression is a statistical method for studying the relationship between several covariates and binary response. In bioinformatics, it is employed to explore disease-associated genes by using the gene expression levels obtained from RNA-seq as covariates. However, it is often difficult to increase the sample size of RNA-seq, especially in the cases of diseases for which clinically inaccessible tissues (e.g. brain) are responsible. In contrast, it is relatively easy to expand the sample size of Transcriptome-wide association study (TWAS), an analytical approach where transcription levels of each gene are estimated solely from the information of individuals' DNA sequence information. This is because the required data of DNA sequence, which is basically the same in all tissues and cells, can be obtained from readily accessible samples (e.g., blood or saliva). Meanwhile, since actual measurements in the responsible tissues are not obtained in TWAS, the predicted values inevitably contain errors, and the extent of these errors can considerably vary across genes.

Motivated by the above problem, we develop a statistical method for integrating data with and without covariate measurement error in logistic regression. Naive application of logistic regression ignoring the measurement error may lead to significant bias in parameter estimates, which is sometimes called an attenuation effect. Thus, we derive a method for correcting this bias based on asymptotic expansion. Simulation results demonstrate that the proposed method reduces bias and attains smaller mean squared error than the naive method.

This is a joint work with Emiko Koyama (RIKEN Center for Brain Science) and Atsushi Takata (RIKEN Center for Brain Science).