別　紙

論 文 の 内 容 の 要 旨

生産・環境生物学専攻　専　攻
平成 21 年度博士課程　入学
モハマド　マニル　ホセイン　モラー
指導教員名 岸野　洋久

Robust Inference and Model Diagnosis of Microarray Data by β-Likelihood
（β-尤度法によるマイクロアレイデータの頑健推定とモデル診断）

1. Introduction

Microarray data enables the high-throughput survey of mRNA expression profiles in genomic level. At the same time, it offers a challenging statistical problem due to the large number of genes surveyed with small sample of sizes. Identification of differentially expressed (DE) genes between two or more user defined groups is an important task to reduce the dimensionality of microarray data. There are several classical and Bayesian or empirical Bayes approaches for identification of DE genes. However, given the complexity of the microarray data, there are no models those can explain the data fully. It is generally difficult to scrutinize the irregular patterns of expression or contaminated genes that are not expected by the models gene by gene.

A statistical framework to detect irregular patterns of expression or contaminated genes and diagnose the model may reduce this difficulty. Inference about deferential expression is a typical objective in analysis of gene expression data. Bayesian approaches have become increasingly popular for detection of DE genes from microarray data. However, most of these approaches are very sensitive to outlier and produces misleading results. Therefore, as an extension of empirical Bayes (EB) procedures, I developed β-EB approaches assuming (i) constant gene-specific variance

and (ii) variable gene-specific variance for the identification of DE genes. Also an attempt is made to extend the β-EB LNN approaches for paired gene expression data analysis. The proposed β-EB approaches are unique parametric approaches because, not only it is robust against outliers, but it also detects contaminating genes and statistically diagnoses gene expression profiles.

To robustify classical EB-approach, I maximize β-likelihood function using EM like algorithm, where the β-likelihood function is induced from the β-divergence. The proposed robust β-EB approaches introduce weight function, which I call β-weight function. The weight of a transcript t is described as a power function of its likelihood, $f^{\beta}(\mathbf{y}_t|\boldsymbol{\theta})$. Genes with low likelihoods have unexpected expression patterns and low weights. By assigning low weights to outliers, the inference becomes robust. The value of β, which controls the balance between the robustness and efficiency, is selected by maximizing the predictive β_0-likelihood by cross-validation. The distribution of the weights is used to scrutinize the irregular patterns of expression and diagnose the model statistically.

2. Outline of the thesis

In chapter 1, I introduced microarray gene expression data, problem of the study and objective of the study in details.

Chapter 2 discusses the microarray technology for generating the gene expression data. This chapter also discuss several classical and Bayesian methods for identification of DE genes, where EBarrays which I call EB in this thesis is one of the most popular approaches for identification of DE genes. However, most of the existing algorithms are not robust against outliers and none of them can detect contaminated genes from high-dimensional gene expression datasets. To overcome this problems, in this thesis, I proposed some modification of classical EB approaches by maximizing β-likelihood function using EM like algorithm for robust statistical inference those are discussed in the subsequent chapters of this thesis.

Chapter 3 introduced β-EB approach for robust identification of DE genes assuming constant gene-specific variance. Numerical simulation showed that the contaminated genes are detected by inspecting the values of the β-weights. In the

absence of contaminated genes, β-EB and the other existing procedures had similar performance. When the data includes contaminated genes, β-EB and not the others were robust.

In chapter 4, I developed β-EB approach for robust identification of DE genes assuming variable gene-specific variance. To see the performance of the estimation and model diagnosis, I conducted two types of simulations. The first simulation compares the performance for different sizes of data and different levels of outliers. The second simulation generates gamma distributed expression profiles, whereas estimation procedures assume log-normal distributions. When the Gamma distribution has the shape parameter < 1, it has a large mass near the value of zero, and will not be approximated by the log-normal distributions. Again, it was shown that the β-EB approach, and not the others, is robust against outliers. By comparing the β-weight distribution with the predicted distribution, it was possible to detect the genes with expression profiles that contradict log-normal distribution.

In chapter 5, I analyzed three sets of real gene expression data (head and neck cancer, lung cancer and Arabidopsis thaliana) using both classical and proposed EB-LNNMV approach. In the analysis of head and neck cancer data, the β-EB approach detected six contaminating genes (LRP8, S100A8, S100A9, TRIM29, CSTA, ACP5) as outliers with the posterior probability of DE > 0.95; the posterior probability for these genes by the classical EB-LNN approach was < 0.5. Inspection of the expression profiles of these genes confirmed the presence of outliers. The classical EB approach had the low power due to over estimation of variances within the groups. In the lung cancer data, the β-weight distribution deviated largely from the predicted distribution, and implied the sign of model misspecification. The analysis of scatter plot showed that this is due to the genes with little expression and the genes with large within variance. By excluding the genes with extremely low expression levels, the β-weight distribution became consistent with the model accompanied by a few outliers. When applied to the eQTL analysis of Arabidopsis thaliana, the β-EB approach gave on average larger numbers of regulated genes compare with classical EB approach. Furthermore, the proposed β-EB approach identified some potential master regulators that were missed by the EB approach. They include markers on a telomeric region of chromosome 4. This region includes three transcription factors one of which is CYC1 (cyclin-dependent protein kinase regulator).

Chapter 6 presents the modification of β-EB approaches for identification of DE genes in the case of paired genes expression data. Simulation results show that the performance of the proposed method is good for identification DE genes from paired observations.

Chapter 7 presents the overall conclusion.

3. Conclusion

In my thesis, I have discussed the robustification of EB approach by β-divergence assuming both constant and variable gene-specific variance. The proposed method reduces to the standard EB approach for $\beta \rightarrow 0$. The performance of the proposed method in a comparison of the classical EB approach, t-test for identification of DE genes investigated using AUC and pAUC in the simulation study. From the simulation results, I observe that the proposed method significantly improves the performance in a comparison of the others in presence of outliers; otherwise, it keeps almost equal performance.

Simulation and real gene expression data analysis results show that the performance of the proposed method much better than the other existing methods in presence of irregular gene expression patterns. Otherwise, it shows almost equal performance.

I extended our proposed β-EB approach for identification of DE genes from the correlated expressions between two user-defined groups. I investigate the performance of this approach using simulated data only. Therefore, I would like to apply this update version to real paired gene expression dataset soon to investigate the performance from the robustness point of view.