Optimal Shrinkage Estimation of Variances With Applications to Microarray Data Analysis

Tiejun TONG and Yuedong WANG

Microarray technology allows a scientist to study genomewide patterns of gene expression. Thousands of individual genes are measured with a relatively small number of replications, which poses challenges to traditional statistical methods. In particular, the gene-specific estimators of variances are not reliable and gene-by-gene tests have low powers. In this article we propose a family of shrinkage estimators for variances raised to a fixed power. We derive optimal shrinkage parameters under both Stein and squared loss functions. Our results show that the standard sample variance is inadmissible under either loss function. We propose several estimators for the optimal shrinkage parameters and investigate their asymptotic properties under two scenarios: large number of replications and large number of genes. We conduct simulations to evaluate the finite sample performance of the data-driven optimal shrinkage estimators and compare them with some existing methods. We construct F-like statistics using these shrinkage variance estimators and apply them to detect differentially expressed genes in a microarray experiment. We also conduct simulations to evaluate performance of these F-like statistics and compare them with some existing methods.

KEY WORDS: F-like statistic; Gene expression data; Inadmissibility; James–Stein shrinkage estimator; Loss function.

1. INTRODUCTION

The development of microarray technology has revolutionized the study of molecular biology and become a standard tool in genomics research. Instead of working on a gene-by-gene basis, microarray technology allows the scientists to view the expression of thousands of genes from an experimental sample simultaneously (Nguyen, Arpat, Wang, and Carroll 2002; Leung and Cavalieri 2003). Due to the cost, it is common that thousands of genes are measured with a small number of replications (Lönnstedt and Speed 2002; Kendziorski, Newton, Lan, and Could 2003). As a consequence, we are faced with a "large G, small n" paradigm, where G is the total number of genes and n is the number of replications. The standard gene-specific estimators of variances are unreliable due to the relatively small number of replications. Consequently, the commonly used statistical methods, such as t test or F test, for detecting differentially expressed genes on a gene-by-gene basis have low powers (Callow, Dudoit, Gong, Speed, and Rubin 2000). On the other hand, the assumption that variances are equal for all genes is unlikely to be true. Thus, tests based on a pooled common variance estimator for all genes are at the risk of generating misleading results (Cui, Hwang, Qiu, Blades, and Churchill 2005).

A number of approaches to improving variance estimation and hypothesis testing have emerged. Kamb and Ramaswami (2001) suggested a simple regression estimation of local variances. Storey and Tibshirani (2003) added a small constant to the gene-specific variance estimators in their SAM t test to stabilize the small variances. Lin, Nadler, Lan, Attie, and Yandell (2003) proposed a data-adapted robust estimator of array error based on a smoothing spline and standardized local median absolute deviation. Jain et al. (2003) proposed a local-pooled-error estimation procedure, which borrows strength from genes in local intensity regions to estimate array error variability. Baldi and Long (2001) proposed a regularized t test by replacing the usual variance estimator with a Bayesian estimator. Lönnstedt and Speed (2002) proposed an empirical Bayes approach that combines information across genes. Kendziorski et al. (2003) extended the empirical Bayes method using hierarchical gamma–gamma and lognormal–normal models.

Cui and Churchill (2003) compared three variance estimators: the gene-specific estimator, the pooled estimator across genes, and the hybrid estimator as the average of the gene-specific and the pooled estimators. Applying the standard James–Stein shrinkage method to log transformed estimates of variances, Cui et al. (2005) proposed a James–Stein type shrinkage estimator for variances (referred to as the CHQBC estimator in the remainder of this article). Compared to some existing tests, they showed that the F test using the James–Stein type variance estimator has the best or nearly the best power to detect differentially expressed genes over a wide range of situations.

The research so far has concentrated on the methodology. Little is known about the theoretical properties of various shrinkage variance estimators. Shrinkage variance estimation has a long history that began with the amazing inadmissibility result discovered by Stein (1964), where the standard sample variance is improved by a shrinkage estimator using information contained in the sample mean. Much research has been done since then (Maatta and Casella 1990; Kubokawa 1999), most of which concerned single variances (Kubokawa 1999), which are not applicable to microarray data analysis because the homogeneity of the variances is unlikely to be true. Some research has been devoted to the shrinkage estimator of a covariance matrix (Kubokawa and Srivastava 2003). However, all these methods require $n > G$ to ensure nonsingularity of the sample covariance matrix. Therefore, these methods break down for microarray data analysis.

We propose new optimal shrinkage estimators in this article. Instead of using information in the sample mean (Stein 1964), we borrow information across variances. We will show that the standard sample variance is inadmissible. Therefore,

Tiejun Tong is Postdoctoral Associate, Department of Epidemiology and Public Health, Yale University, New Haven, CT 06520 (E-mail: *tiejun.tong@yale.edu*). Yuedong Wang is Professor, Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106 (E-mail: *yuedong@pstat.ucsb.edu*). This research was supported by National Institutes of Health grant R01 GM58533. The authors thank the joint editor, the associate editor, and two referees for their constructive comments and suggestions that have led to a substantial improvement in the article. The authors also thank Xiangqin Cui and Gary Churchill in the Jackson Laboratory for providing the data and answering many technical questions about the MAANOVA package.

Journal of the American Statistical Association
March 2007, Vol. 102, No. 477, Theory and Methods
DOI 10.1198/016214506000001266