Variance Estimation in the Analysis of Microarray Data

Yuedong Wang
Department of Statistics and Applied Probability
University of California, Santa Barbara, California 93106
yuedong@pstat.ucsb.edu

Yanyuan Ma
Department of Statistics
Texas A&M University, College Station TX 77843-3143
ma@stat.tamu.edu

and Raymond J. Carroll
Department of Statistics
Texas A&M University, College Station TX 77843-3143
carroll@stat.tamu.edu

Abstract

In microarrays analysis, one main problem is that conventional estimates of the variances are unreliable due to the small number of replications. It is commonly observed that the variance increases proportionally with the intensity level, leading to modeling variance as a function of mean through the constant coefficient of variation model and the quadratic variance-mean model. Because the means are unknown and estimated with few degrees of freedom, naive methods that use the sample mean in place of the true mean are biased. We propose three methods for overcoming this bias. The first two are variations of a heteroscedastic-simulation-extrapolation estimator, the third is based on semiparametric information calculations. Theory and Simulations show the power of our methods and their lack of bias compared to the naive method. The methodology is illustrated using microarray data from leukemia patients.

Some Key Words: Heteroscedasticity, Measurement error, Microarray, Semiparametric methods, Simulation-extrapolation, Variance function estimation.

1 Introduction

Microarrays are one of the most widely used high-throughput technologies, enabling scientists to simultaneously measure the expression of thousands of genes (Nguyen, Arpat, Wang & Carroll 2002, Leung & Cavalieri 2003). A microarray experiment typically involves a large number of genes and a relatively small number of replications. This new paradigm presents many challenges to standard statistical methods. For example, the standard t-test for detecting differentially expressed genes under two experimental conditions usually has low power (Callow, Dudoit, Gong, Speed & Rubin 2000, Cui, Hwang, Qiu, Blades & Churchill 2005).

One of the main problems is that conventional estimates of the variances required in the t-statistic and other statistics are unreliable due to the small number of replications. Various methods

have been proposed in the literature to overcome this lack of degrees of freedom problem (Rocke & Durbin 2001, Kamb & Ramaswami 2001, Huang & Pan 2002, Storey & Tibshirani 2003, Lin, Nadler, Attie & Yandell 2003, Jain, Thatte, Braciale, Ley, O'Connell & Lee 2003, Strimmer 2003, Tong & Wang 2006). A key idea for getting better estimates of variances is to borrow information from different genes with similar variances. It is commonly observed that the variance increases proportionally with the intensity level, which has led many authors to assume that the variance is a function of the mean (Chen, Dougherty & Bittner 1997, Rocke & Durbin 2001, Huang & Pan 2002). Chen et al. (1997), Rocke & Durbin (2001), Chen, Kamat, Dougherty, Bittner, Meltzer & Trent (2002) and Weng, Dai, Zhan, He, Stepaniants & Bassett (2006) modeled the variance-mean function parametrically while Kamb & Ramaswami (2001), Huang & Pan (2002), Lin et al. (2003) and Jain et al. (2003) modeled it nonparametrically. We will limit ourself to parametric variance-mean models in this article. Specifically, for simplicity and applicability in microarray data analysis, we will concentrate on two models: the constant coefficient of variation model proposed by Chen et al. (1997) and the quadratic variance-mean model proposed by Rocke & Durbin (2001) and Chen et al. (2002). Of course, our results can be generalized to other parametric models, but since the two mentioned about are often used, we confine our attention to them.

Strimmer (2003) fitted the quadratic variance-mean model using quasi-likelihood. He estimated parameters in the variance function together with all mean parameters for each gene. Since the number of genes is large, it is likely that the estimates of variance parameters are inconsistent, i.e., this is a Neyman-Scott type problem. Strimmer found that the variance parameters are underestimated in his simulations. An alternative approach which could lead to consistent estimates of variance parameters is to fit a variance-mean model using reduced data consisting of sample means and variances (Huang & Pan 2002). However, as we will illustrate in this article, due to sampling error that has a similar effect here as measurement errors, which has not been noted in the literature, naive estimates based on sample means and variances are inconsistent. We will also show that the well-known simulation extrapolation (SIMEX) method fails to correct biases in some estimators and propose new consistent estimators.

Our key insight into this problem is that technically it is closely related to a measurement error problem (Carroll, Ruppert, Stefanski & Crainiceanu 2006) where the measurement error has nonconstant variance and the structure of the variance function is of interest. Thus it is amenable to analyses similar to measurement error models. However, because of the special structure of the problem, where independence between the measurement error and regression model as in classical measurement error model fails, and the fact that it is the variance function itself that is of interest, direct application of measurement error methods typically does not work. This requires new methods that do not exist in the standard measurement error literature.

In this paper, we propose two methods for attacking the problem.

- The first is a novel modification of the SIMEX method, which we call the permutation SIMEX. The key notion is that the ordinary SIMEX method requires that the responses and the additional noise added in a part of the algorithm be independent. In our problem, this independence does not hold. Our method breaks this connection between the response and the noise, thus allowing the possibility of consistent estimation that classical SIMEX is not able to obtain.

- The second approach is based on our insight of casting the problem in a semiparametric framework while treating the unobservable variable distribution as a nuisance parameter.

We employ a projection approach to achieve consistency without making any distributional assumptions about the mean gene expression.

2 Results

The central model of interest arising from microarray data analysis has the form

$$Y_{i,j} = X_i + g^{1/2}(X_i; \boldsymbol{\theta})\epsilon_{i,j}, \quad i = 1, \cdots, n; \quad j = 1, \cdots, m, \tag{1}$$

where $Y_{i,j}$ is the j^{th} replicate of observed expression level of gene i, X_i is the expected expression level of gene i, $\epsilon_{i,j}$ are independent random errors with mean 0, variance 1 and at least finite fourth moments, and $\boldsymbol{\theta}$ is a d-dimensional parameter vector. For convenience, throughout the paper we assume that $\epsilon_{i,j}$ is a standard normal random variable. As in any SIMEX-type method, strictly speaking this normality is required, although it is well-known that the methods are robust to modest departures from normality (Carroll et al, 2006, p. 101). The semiparametric methods can be applied for any distribution. Our goal is to estimate $\boldsymbol{\theta}$ in the variance function $g(\cdot)$ from the observations Y_{ij}'s, for $i = 1, \ldots, n, j = 1, \ldots, m$.

The most popular parametric models for the variance function in the microarray data analysis literature include the constant coefficient of variation model and the quadratic variance-mean model. The constant coefficient of variation model has the form

$$g(x; \theta) = \theta x^2, \quad \theta \geq 0. \tag{2}$$

Chen et al. (1997) assumed this model for cDNA microarray data. While it is adequate for genes with high expression levels, it is inaccurate when the signal is weak in comparison to the background. To overcome this problem, Rocke & Durbin (2001), Chen et al. (2002) and Strimmer (2003) considered the following quadratic model:

$$g(x; \boldsymbol{\theta}) = \alpha + \beta x^2, \quad \alpha \geq 0, \quad \beta \geq 0, \tag{3}$$

where $\boldsymbol{\theta} = (\alpha, \beta)$. For ease of exposition, we assume that the background (stray) signal has been removed. One may estimate the background signal by including a linear term in model (3) (Strimmer 2003).

2.1 The Permutation SIMEX Estimator

The fact that S_i and $W_{b,i}(\zeta)$ are constructed from the same repeated measures $Y_{i,j}$'s can cause perfectly plausible estimators to fail to extrapolate correctly because of the induced correlation of the response and the measurement errors. We now describe a method that guarantees correct extrapolation, in the sense that the limiting value as first $n \to \infty$ and then $\zeta = -1$ is the correct population-level quantity.

The main idea is to "break" the connection between the response and the measurement errors, and force nondifferential error, thus placing the estimator within the context of standard heteroscedastic-SIMEX. The method requires that $m \geq 3$.

2.2 The Semiparametric Estimator

The insight of viewing the unobservable variable X_i as latent allows us to treat the problem in the semiparametric framework. The choice of using a projection approach instead of estimating the latent variable distribution, while still achieving consistency, makes the approach very appealing. As far as we know, despite the fact that general semiparametric methodology is well developed, no consistent estimator is known for this specific problem.

To facilitate the computation of multi-dimensional integration, we consider here a slightly more general model $Y_{i,j} = X_i + a_j g^{1/2}(X_i; \boldsymbol{\theta})\epsilon_{i,j}$. The only difference between this model and the one in (1) is the inclusion of the known constants $a_j, j = 1, \ldots, m$. The original model (1) corresponds to $a_j = 1$. The need of such generalization will become evident when we look into the implementation in Section **??**. The probability density function (pdf) of a single observation $\mathbf{Y}_i = (Y_{i,1}, \ldots, Y_{i,m})^T$ is

$$p_{\mathbf{Y}}(\mathbf{Y}_i, \boldsymbol{\theta}, \eta) = C \int \eta(X_i)\{g(X_i; \boldsymbol{\theta})\}^{-m/2} \exp\left\{ -\frac{1}{g(X_i; \boldsymbol{\theta})} \sum_{j=1}^{m} \frac{(Y_{i,j} - X_i)^2}{2a_j^2} \right\} d\mu(X_i),$$

where C is a constant and $\eta(X_i)$ represents the unspecified density function of the latent variable X_i. The problem of estimation of $\boldsymbol{\theta}$ is thus a semiparametric estimation problem. We proceed to construct a class of semiparametric estimator of $\boldsymbol{\theta}$ through deriving its efficient influence function. The efficient influence function contains the unknown nuisance parameter $\eta(\cdot)$, the estimation of which is difficult. In line with several related techniques (Tsiatis & Ma 2004, Ma, Genton & Tsiatis 2005), we avoid estimating $\eta(\cdot)$, and argue instead that various possibly misspecified $\eta^*(\cdot)$ can be plugged into the result estimating equation to obtain a class of consistent estimators. When $\eta^*(\cdot)$ happens to be the truth, denoted by $\eta_0(\cdot)$, then the resulting estimator is optimal in terms of its asymptotic efficiency.

3 Conclusion

The key insights of this article are that the naive approach of ignoring sampling error will lead to inconsistent estimates, and the well-known heteroscedastic-SIMEX approach to dealing with the measurement error should be applied with caution, especially outside the constant coefficient of variation model. Two parametric variance-mean models used in microarray data analysis, the constant coefficient of variation model and the quadratic variance-mean model, are used to illustrate these insights. We believe that the inconsistency problems associated with the naive and direct SIMEX estimators persist for general models and the proposed permutation SIMEX and semiparametric methods work for general models.

The key to our analysis of SIMEX-type methods was to note that direct application of standard heteroscedastic-SIMEX will not generally work because of an induced differential measurement error. Our permutation SIMEX approach avoids this problem, forcing nondifferential error, and in all cases considered equals or vastly outperforms ordinary heteroscedastic-SIMEX. The key to our semiparametric method was to note that this is indeed a measurement error problem, and to realize that grouping observations can lead to great gains in computationally efficiency. Both the theoretical derivation and simulation studies demonstrated the satisfactory performance of our two methods in terms of asymptotic consistency and valid inference.

One important future research topic is to evaluate the impact of the proposed methods on microarray data analysis and compare them with alternative methods such as VarMixt (Delmar, Robin & Daudin 2005) and data-driven Haar-Fisz (Motakis, Nason, Fryzlewicz & Rutter 2006).

References

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P. & Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in hdl-deficient mice, *Genome Research* **10**: 2022–2029.

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, 2nd ed.*, Chapman & Hall, New York.

Chen, Y., Dougherty, E. R. & Bittner, M. L. (1997). Ratio-based decisions and the quantitative analysis of cdna microarray images, *Journal of Biomedical Optics* **2**: 364–374.

Chen, Y., Kamat, V., Dougherty, E. R., Bittner, M. L., Meltzer, P. S. & Trent, J. M. (2002). Ratio statistics of gene expression levels and applications to microarray data analysis, *Bioinformatics* **18**: 1207–1215.

Cui, X., Hwang, J. T. G., Qiu, J., Blades, N. J. & Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates, *Biostatistics* **6**: 59–75.

Delmar, P., Robin, S. & Daudin, J. J. (2005). Varmixt: efficient variance modelling for the differential analysis of replicated gene expression data, *Bioinformatics* **21**: 502–508.

Huang, X. & Pan, W. (2002). Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays, *Funct Integr Genomics* **2**: 126–133.

Jain, N., Thatte, J., Braciale, T., Ley, K., O'Connell, M. & Lee, J. (2003). Local-pooled error test for identifying differentially expressed genes with a small number of replicated microarrays, *Bioinformatics* **19**: 1945–1951.

Kamb, A. & Ramaswami, A. (2001). A simple method for statistical analysis of intensity differences in microarray-derived gene expression data, *BMC Biotechnol.* pp. 1–8.

Leung, Y. & Cavalieri, D. (2003). Fundamentals of cDNA microarray data analysis, *TRENDS in Genetics* **11**: 649–659.

Lin, Y., Nadler, S. T., Attie, A. D. & Yandell, B. S. (2003). Adaptive gene picking with microarray data: detecting important low abundance signals. in Parmigiani, G., Garrett, E. S., Irizarry, R. A. and Zeger, S. L. (ed.), *The Analysis of Gene Expression Data: Methods and Software* . New York: Springer.

Ma, Y., Genton, M. G. & Tsiatis, A. A. (2005). Locally efficient semiparametric estimators for generalized skew-elliptical distributions, *Journal of the American Statistical Association* **100**: 980–989.

Motakis, E. S., Nason, G. P., Fryzlewicz, P. & Rutter, G. A. (2006). Variance stabilization and normalization for one-color microarray data using a data-driven multiscale approach, *Bioinformatics* **22**: 2547–2553.

Nguyen, D. V., Arpat, A. B., Wang, N. & Carroll, R. J. (2002). DNA microarray experiments: Biological and technological aspects, *Biometrics* **58**: 701–717.

Rocke, D. M. & Durbin, B. (2001). A model for measurement error for gene expression arrays, *Journal of Computational Biology* **8**: 557–569.

Storey, J. & Tibshirani, R. (2003). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. in Parmigiani, G., Garrett, E. S., Irizarry, R. A. and Zeger, S. L. (ed.), *The Analysis of Gene Expression Data: Methods and Software* . New York: Springer.

Strimmer, K. (2003). Modeling gene expression measurement error: a quasi-likelihood approach, *BMC Bioinformatics* **4**.

Tong, T. & Wang, Y. (2006). Optimal shrinkage estimation of variances with applications to microarray data analysis, to appear in *Journal of the American Statistical Association* .

Tsiatis, A. A. & Ma, Y. (2004). Locally efficient semiparametric estimators for functional measurement error models, *Biometrika* **91**: 835–848.

Weng, L., Dai, H., Zhan, Y., He, Y., Stepaniants, S. B. & Bassett, D. E. (2006). Rosetta error model for gene expression analysis, *Bioinformatics* pp. 1111–1121.