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 Institut de Statistique, Université de Neuchâtel
Pierra à Mazel 7, 2000 Neuchâtel, Switzerland
Yanyuan.Ma@unine.ch

and Raymond J. Carroll
Department of Statistics
Texas A&M University, College Station TX 77843-3143
carroll@stat.tamu.edu
July 23, 2008

Summary

Microarrays are one of the most widely used high-throughput technologies. One of the main problems in the area 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. In this context, 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. Here we concentrate on estimation of the variance as a function of an unknown mean in two models: 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 generally biased because of the errors-in-variables phenomenon. In this paper we propose three methods for overcoming this bias. The first two are variations on the theme of the so-called heteroscedastic-SIMEX estimator, modified to consistently estimate the variance function. The third class of estimators is entirely different, being based on semiparametric information calculations. Simulations show the power of our methods and their lack of bias compared to the naive method that ignores the measurement error. The methodology is illustrated using microarray data from leukemia patients.

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

Short Title: 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 and Carroll 2002, Leung and 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 and Rubin 2000, Cui, Hwang, Qiu, Blades and 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 and Durbin 2001, Kamb and Ramaswami 2001, Huang and Pan 2002, Storey and Tibshirani 2003, Lin, Nadler, Attie and Yandell 2003, Jain, Thatte, Braciale, Ley, O'Connell and Lee 2003, Strimmer 2003, Tong and 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 and Bittner 1997, Rocke and Durbin 2001, Huang and Pan 2002). Chen et al. (1997), Rocke and Durbin (2001), Chen, Kamat, Dougherty, Bittner, Meltzer and Trent (2002) and Weng, Dai, Zhan, He, Stepaniants and Bassett (2006) modeled the variance-mean function parametrically while Kamb and Ramaswami (2001), Huang and 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 and 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 un-

derestimated 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 and 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 and 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.

We consider an asymptotic approach for increasing number of genes and fixed number of replications. Section 2 introduces the model and briefly describes a moment estimator and a regression estimator for the constant coefficient of variation model and the quadratic variation model, respectively. We show that the naive estimators are inconsistent. Section 3 applies a simulation extrapolation (SIMEX) approach to the moment estimators and shows that the resulting estimators

are consistent. Section 4 illustrates that in general, SIMEX is prone to be implemented improperly, because of the special structure of our problem. A novel modified SIMEX-type methodology that applies to all parametric models is described. Because it is based on a permutation-type philosophy, it is termed permutation SIMEX. Section 5 takes an entirely different approach, and casts this problem within the context of semiparametric models (Bickel, Klaassen, Ritov and Wellner 1993). We show how to construct general estimators that are consistent and have local semiparametric efficiency. We apply the methods to a data example in Section 6 and conduct simulation studies in Section 7. Discussion and concluding remarks are given in Section 8. All the technical derivations are provided in an Appendix. Derivations that are largely algebraic in nature are included in Supplemental Materials available at http://www.pstat.ucsb.edu/faculty/yuedong. Computer code is also available at the same web site.

2 The Model

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 and 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).

For simplicity, we use the notations $g(x;\theta)$ and $g(x)$ interchangeably. Let $\overline{Y}_{i,\cdot}$ be the sample mean for the i^{th} unit, and let S_i be the sample variance based on $Y_{i,j}$. In the hypothetical situation when X is observed, two simple consistent estimators can be obtained using either the method of moments or least squares (LS). For the constant coefficient of variation model (2), the two estimators have the form

$$\widehat{\theta}_M = \frac{n^{-1}\sum_{i=1}^n S_i}{n^{-1}\sum_{i=1}^n X_i^2}; \tag{4}$$

$$\text{and} \qquad \widehat{\theta}_R \triangleq \arg\min_{\theta}\left\{ n^{-1}\sum_{i=1}^n (S_i - \theta X_i^2)^2 \right\} = \frac{n^{-1}\sum_{i=1}^n X_i^2 S_i}{n^{-1}\sum_{i=1}^n X_i^4}, \tag{5}$$

while for the quadratic model (3), they are respectively given as

$$\begin{aligned} \widehat{\alpha}_M &= n^{-1}\sum_{i=1}^n S_i - \widehat{\beta}_M n^{-1}\sum_{i=1}^n X_i^2, \\ \widehat{\beta}_M &= \sqrt{\frac{\frac{m-1}{m+1} n^{-1}\sum_{i=1}^n S_i^2 - (n^{-1}\sum_{i=1}^n S_i)^2}{n^{-1}\sum_{i=1}^n X_i^4 - \{n^{-1}\sum_{i=1}^n X_i^2\}^2}}; \end{aligned} \tag{6}$$

$$\begin{aligned} \text{and} \qquad \widehat{\alpha}_R &= n^{-1}\sum_{i=1}^n S_i - \widehat{\beta}_R n^{-1}\sum_{i=1}^n X_i^2, \\ \widehat{\beta}_R &= \frac{n^{-1}\sum_{i=1}^n X_i^2 S_i - (n^{-1}\sum_{i=1}^n X_i^2)(n^{-1}\sum_{i=1}^n S_i)}{n^{-1}\sum_{i=1}^n X_i^4 - \{n^{-1}\sum_{i=1}^n X_i^2\}^2}. \end{aligned} \tag{7}$$

The LS estimators in (7) are minimizers of $n^{-1}\sum_{i=1}^n (S_i - \alpha - \beta X_i^2)^2$. For simplicity, LS instead of weighted LS is used. The moment estimator $\widehat{\theta}_M$ is derived from the equation matching the first moment $\theta\sum_{i=1}^n X_i^2 = \sum_{i=1}^n S_i$. The moment estimators $\widehat{\alpha}_M$ and $\widehat{\beta}_M$ are derived from the following equations matching the first two moments:

$$\begin{aligned} \alpha + \beta n^{-1}\sum_{i=1}^n X_i^2 &= n^{-1}\sum_{i=1}^n S_i, \\ \alpha^2 + 2\alpha\beta n^{-1}\sum_{i=1}^n X_i^2 + \beta^2 n^{-1}\sum_{i=1}^n X_i^4 &= \frac{m-1}{m+1} n^{-1}\sum_{i=1}^n S_i^2. \end{aligned} \tag{8}$$

Note that the normality assumption was used in the derivation of (8). We take the positive square root for β since $\beta \geq 0$.

Since $\overline{Y}_{i,\cdot}$ is an unbiased estimator of X_i, a naive approach in the absence of X_i's is to replace X_i in (4)-(7) by $\overline{Y}_{i,\cdot}$. Unfortunately, this approach ignores the sampling error in $\overline{Y}_{i,\cdot}$, and leads to inconsistent estimates in general. See Lemmas 1 and 4 in the Appendix Sections A.1 and A.2 for detailed calculations. Asymptotically, the parameters θ in model (2) and β in model (3) are under-estimated, resulting in the classic problem of attenuation to the null. Throughout this paper, asymptotics are based on $n \to \infty$ with a fixed m.

3 The SIMEX Moment Estimator

Models described in (1) with X_i unobserved are latent variable models. They can also be viewed as heteroscedastic measurement error models, because by their very nature the $Y_{i,j}$ are error-prone unbiased measures of X_i with non-constant variation. This viewpoint enables us to adopt a SIMEX method developed in the heteroscedastic measurement error model framework, see Devanarayan and Stefanski (2002). The method requires that we specify a method for parameter estimation in the case that X_i is observed. When the method of moments is specified, the resulting SIMEX algorithm is the following:

1. Generate $Z_{b,i,j} \overset{iid}{\sim} N(0,1)$, $i = 1, \cdots, n$, $j = 1, \cdots, m$, $b = 1, \cdots, B$. Let

$$c_{b,i,j} = \frac{Z_{b,i,j} - \overline{Z}_{b,i,\cdot}}{\sqrt{\sum_{j=1}^{m}(Z_{b,i,j} - \overline{Z}_{b,i,\cdot})^2}}.$$

2. For $i = 1, \cdots, n$, $j = 1, \cdots, m$, $b = 1, \cdots, B$, let $W_{i,j} = Y_{i,j}$ and

$$W_{b,i}(\zeta) \quad = \quad \overline{W}_{i,\cdot} + (\frac{\zeta}{m})^{1/2} \sum_{j=1}^{m} c_{b,i,j} W_{i,j}.$$

   Then $\mathrm{E}\{W_{b,i}(\zeta)|X_i\} = X_i$ and $\mathrm{var}\{W_{b,i}(\zeta)|X_i\} = \{(1+\zeta)/m\}g(X_i)$.

3. Estimate $\boldsymbol{\theta}$ by replacing X_i in (4) and (6) by $W_{b,i}(\zeta)$ for each b and then average over b.

4. Extrapolate back to $\zeta = -1$.

In general, the essential idea of a SIMEX-type method is to add via simulation (the SIM step) increasing amounts of measurement error to understand how measurement error affects a parameter

estimate, and then to extrapolate (the EX step) back to the case of no measurement error. Steps 1, 2 and 3 above are the SIM step for heteroscedastic models, although see Section 4 for a subtlety. Note that as ζ increases, since $\text{var}\{W_{b,i}(\zeta)|X_i\} = \{(1+\zeta)/m\}g(X_i)$, the measurement error also increases, and in this sense the $W_{b,i}(\zeta)$ fulfill the requirement of adding noise. Also note that when $\zeta = -1$, $\text{var}\{W_{b,i}(\zeta)|X_i\} = 0$, and hence extrapolating back to $\zeta = -1$ is a means to obtain an estimator that avoids bias.

The resulting estimators for models (2) and (3) as a function of ζ are

$$\begin{aligned} \widehat{\theta}_{S.M}(\zeta) &= B^{-1}\sum_{b=1}^{B}\frac{n^{-1}\sum_{i=1}^{n}S_i}{n^{-1}\sum_{i=1}^{n}W_{b,i}^2(\zeta)}; \\ \text{and} \qquad \widehat{\alpha}_{S.M}(\zeta) &= n^{-1}\sum_{i=1}^{n}S_i - \widehat{\beta}_{S.M}(\zeta)B^{-1}\sum_{b=1}^{B}n^{-1}\sum_{i=1}^{n}W_{b,i}^2(\zeta), \\ \widehat{\beta}_{S.M}(\zeta) &= B^{-1}\sum_{b=1}^{B}\sqrt{\frac{\frac{m-1}{m+1}n^{-1}\sum_{i=1}^{n}S_i^2 - (n^{-1}\sum_{i=1}^{n}S_i)^2}{n^{-1}\sum_{i=1}^{n}W_{b,i}^4(\zeta) - \{n^{-1}\sum_{i=1}^{n}W_{b,i}^2(\zeta)\}^2}}, \end{aligned}$$

respectively for any fixed extrapolant point ζ.

Theorem 1. The SIMEX approach leads to consistent moment estimators. Specifically, $\widehat{\theta}_{S.M}(\zeta) \xrightarrow{p} s_1(\zeta)$, $\widehat{\alpha}_{S.M}(\zeta) \xrightarrow{p} s_2(\zeta)$, $\widehat{\beta}_{S.M}(\zeta) \xrightarrow{p} s_3(\zeta)$ as $n \to \infty$, and $s_1(-1) = \theta$, $s_2(-1) = \alpha$, $s_3(-1) = \beta$. In addition, for any smooth extrapolant function including the correct one is used, the SIMEX moment estimators are asymptotically normally distributed.

[Proof] The actual forms of s_k's can be found in Sections A.1, A.2 and A.3 in the Appendix. The proofs are in the Supplemental Materials.

Remark 1. In Section A.3, we provide an explicit analysis of the SIMEX moment estimator $\widehat{\theta}_{S,M}(\zeta)$ for the constant coefficient of variation model (2). Similar analysis can be performed for all other variance models. In practice however, the bootstrap could be more effective and straightforward albeit computationally expensive, see Carroll et al. (2006) and Section 6 for details.

4 The Permutation SIMEX Estimator

4.1 A Subtlety and Problems with the SIMEX Approach

There is a subtlety that makes Theorem 1 rather surprising. In general, SIMEX-type methods and indeed most measurement error methods require nondifferential measurement error, i.e., the

measurement error is independent of the response. However, this is not the case here: the "response" S_i can be shown to be not independent of $W_{b,i}(\zeta)$, and hence the measurement error in the SIMEX steps is differential. This makes Theorem 1 very unexpected: the measurement error is differential and yet SIMEX works for method of moments.

As it turns out, asymptotic validity of the SIMEX method described in the previous section is not a general phenomenon, and it fails for the regression estimators. Consider the constant coefficient of variation model (2) and the regression through the origin estimator (5). If X_i were observable, then $n^{-1}\sum_i X_i^2 S_i / n^{-1}\sum_i X_i^4$ is a consistent estimator of θ. However, if one uses heteroscedastic-SIMEX and replaces X_i by $W_{b,i}(\zeta)$, the limiting value as $n \to \infty$ for any ζ is (Lemma 2 in the Appendix)

$$\theta \frac{1 + \theta\left\{(1+\zeta)/m + 2\zeta/m(m-1)\right\}}{1 + 6\theta(1+\zeta)/m + 3\theta^2(1+\zeta)^2/m^2},$$

and does not extrapolate to θ when $\zeta = -1$. Similarly, the SIMEX approach fails to correct biases in the regression estimators for the quadratic model (3) (Lemma 5 in the Appendix) as well.

In summary, the usual heteroscedastic-SIMEX approach is not a general prescription for this problem, and we need new methods.

4.2 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 $m \geq 3$. The algorithm is the following, where the construction of $W_{b,i}^{(j)}(\zeta)$ is based on all observations except $Y_{i,j}$ and then its deviation $S_i^{(j)}$ is measured against the $Y_{i,j}$, as in step 1b.

1. Do $j = 1, \cdots, m$,

(a) Generate $Z_{b,i,k} \overset{iid}{\sim} N(0,1)$, $i = 1, \cdots, n$, $k = 1, \cdots, m-1$, $b = 1, \cdots, B$. Let

$$c_{b,i,k}^{(j)} = \frac{Z_{b,i,k} - \overline{Z}_{b,i,\cdot}}{\sqrt{\sum_{k=1}^{m-1}(Z_{b,i,k} - \overline{Z}_{b,i,\cdot})^2}}.$$

(b) For $i = 1, \cdots, n$, $k = 1, \cdots, m-1$, $b = 1, \cdots, B$, let

$$\begin{aligned} W_{i,k}^{(j)} &= \begin{cases} Y_{i,k} & 1 \le k \le j-1, \\ Y_{i,k+1} & j \le k \le m-1, \end{cases} \\ W_{b,i}^{(j)}(\zeta) &= \overline{W}_{i,\cdot}^{(j)} + (\frac{\zeta}{m-1})^{1/2} \sum_{k=1}^{m-1} c_{b,i,k}^{(j)} W_{i,k}^{(j)}, \\ S_i^{(j)} &= \{Y_{i,j} - W_{b,i}^{(j)}(\zeta)\}^2. \end{aligned}$$

Then $\mathrm{E}\{W_{b,i}^{(j)}(\zeta)|X_i\} = X_i$ and $\mathrm{var}\{W_{b,i}^{(j)}(\zeta)|X_i\} = \frac{1+\zeta}{m-1} g(X_i)$.

2. Note that, by construction, $Y_{i,j}$ and $W_{b,i}^{(j)}(\zeta)$ are independent, and hence the measurement error in $W_{b,i}^{(j)}(\zeta)$ as a predictor of $Y_{i,j}$ is nondifferential. It is this fact that makes permutation SIMEX work. Hence, estimate $\boldsymbol{\theta}$ by replacing S_i and X_i in (4)-(7) by $S_i^{(j)}$ and $W_{b,i}^{(j)}(\zeta)$, respectively, for each combination of j and b, and then average over all j and b.

3. Extrapolate to $\zeta = -1$.

Remark 2. An alternative approach when $m \ge 4$ is to split m replications into two parts with at least two replications in each part: one part for computing sample variances (unbiased and independent of simulated samples from SIMEX) and one part for the SIMEX procedure.

Because we have forced nondifferential measurement error by construction, the estimators in step 2 can be either the moment estimator or the regression estimator. For any ζ and the constant coefficient of variation model (2), the moments and regression estimators are respectively,

$$\begin{aligned} \widehat{\theta}_{PS.M}(\zeta) &= B^{-1} \sum_{b=1}^{B} m^{-1} \sum_{j=1}^{m} \frac{n^{-1} \sum_{i=1}^{n} S_i^{(j)}}{n^{-1} \sum_{i=1}^{n} \{W_{b,i}^{(j)}(\zeta)\}^2}; \\ \widehat{\theta}_{PS.R}(\zeta) &= B^{-1} \sum_{b=1}^{B} m^{-1} \sum_{j=1}^{m} \frac{n^{-1} \sum_{i=1}^{n} \{W_{b,i}^{(j)}(\zeta)\}^2 S_i^{(j)}}{n^{-1} \sum_{i=1}^{n} \{W_{b,i}^{(j)}(\zeta)\}^4}. \end{aligned}$$

For the quadratic model (3), the moment and regression estimators are, respectively,

$$\begin{aligned}
\widehat{\alpha}_{PS.M}(\zeta) &= B^{-1}\sum_{b=1}^{B} m^{-1}\sum_{j=1}^{m}\left\{ n^{-1}\sum_{i=1}^{n} S_i^{(j)} - \widehat{\beta}_{PS.M}(\zeta) n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^2 \right\}, \\
\widehat{\beta}_{PS.M}(\zeta) &= B^{-1}\sum_{b=1}^{B} m^{-1}\sum_{j=1}^{m}\sqrt{\frac{\frac{1}{3n}\sum_{i=1}^{n}(S_i^{(j)})^2 - (n^{-1}\sum_{i=1}^{n} S_i^{(j)})^2}{n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^4 - [n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^2]^2}}; \\
\widehat{\alpha}_{PS.R}(\zeta) &= B^{-1}\sum_{b=1}^{B} m^{-1}\sum_{j=1}^{m}\left\{ n^{-1}\sum_{i=1}^{n} S_i^{(j)} - \widehat{\beta}_{PS.R} n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^2 \right\}, \\
\widehat{\beta}_{PS.R}(\zeta) &= B^{-1}\sum_{b=1}^{B} m^{-1}\sum_{j=1}^{m}\frac{n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^2 S_i^{(j)} - [n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^2](n^{-1}\sum_{i=1}^{n} S_i^{(j)})}{n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^4 - [n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^2]^2}.
\end{aligned}$$

Theorem 2. The permutation SIMEX approach leads to consistent moment and regression estimators. Specifically, $\widehat{\theta}_{PS.M}(\zeta) \xrightarrow{p} s_4(\zeta)$, $\widehat{\alpha}_{PS.M}(\zeta) \xrightarrow{p} s_5(\zeta)$, $\widehat{\beta}_{PS.M}(\zeta) \xrightarrow{p} s_6(\zeta)$, $\widehat{\theta}_{PS.R}(\zeta) \xrightarrow{p} s_7(\zeta)$, $\widehat{\alpha}_{PS.R}(\zeta) \xrightarrow{p} s_8(\zeta)$, $\widehat{\beta}_{PS.R}(\zeta) \xrightarrow{p} s_9(\zeta)$ as $n \to \infty$, and $s_4(-1) = s_7(-1) = \theta$, $s_5(-1) = s_8(-1) = \alpha$, $s_6(-1) = s_9(-1) = \beta$.

[Proof] The actual forms of s_k's can be found in the Appendix in Sections A.1 and A.2. The proofs are in the Supplemental Materials.

5 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.

5.1 Method Development

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 5.2. 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 and Ma 2004, Ma, Genton and 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.

The approach we take to derive the influence function is geometric. Consider the Hilbert space $\mathcal{H}$ of all the mean zero functions of $\mathbf{Y}$ with finite variance, where the inner product of $\mathcal{H}$ is defined as the covariance between two functions. Here all the expectations in $\mathcal{H}$ are calculated under the true distribution of $\mathbf{Y}$. We decompose $\mathcal{H}$ into a nuisance tangent space Λ and its orthogonal complement $\Lambda^{\perp}$, so that each function in $\Lambda^{\perp}$ corresponds to an influence function (Bickel et al. 1993, Tsiatis 2006). The efficient influence function can be calculated via orthogonal projection of the score function of $p_{\mathbf{Y}}(\mathbf{Y}_i, \boldsymbol{\theta}, \eta)$ with respect to $\boldsymbol{\theta}$. In the Appendix, we calculate the projection to be

$$S_{eff} = S_{\boldsymbol{\theta}}(\mathbf{Y}) - E\{f(X)|\mathbf{Y}\}, \tag{9}$$

where $S_{\boldsymbol{\theta}}(\mathbf{Y}, \boldsymbol{\theta}, \eta) = \partial \log p_{\mathbf{Y}}(\mathbf{Y}, \boldsymbol{\theta}, \eta)/\partial\boldsymbol{\theta}$ is the score function, and $f(X)$ satisfies

$$E(S_\theta|X) = E[E\{f(X)|\mathbf{Y}\}|X]. \tag{10}$$

Estimation based on the form of S_{eff} given in (9) and (10) is not realistic, since it depends on $\eta(X)$, which itself is unknown and notoriously difficult to estimate. However, the structure of the estimator allows us to plug in an "arbitrary" model $\eta^*(X)$ into the computation and the consistency will be retained. Intuitively, this is because $E(S_{eff}) = E\{E(S_{eff}|X)\}$, while $E(S_{eff}|X) = 0$ is guaranteed by our operation in (10), whether or not computed under a true η.

The algorithm for the semiparametric estimator is the following:

1. Propose a distribution model for the latent variable X_i, say $\eta^*(X)$.

2. Solve for $f(X, \boldsymbol{\theta})$ from the equation

$$E\{S^*_{\boldsymbol{\theta}}(\mathbf{Y})|X\} = E[E^*\{f(X, \boldsymbol{\theta})|\mathbf{Y}\}|X], \tag{11}$$

where $S^*_{\boldsymbol{\theta}}(\mathbf{Y}) = E^*\{\partial \log p_{\mathbf{Y}|X}(\mathbf{Y}|X, \boldsymbol{\theta})/\partial\boldsymbol{\theta}|\mathbf{Y}\}$, E^* represents the expectation calculated under η^*.

3. Form the estimating equation

$$\sum_{i=1}^{n} S^*_{eff}(\mathbf{Y}_i, \boldsymbol{\theta}) = 0 \tag{12}$$

where $S^*_{eff} = S^*_{\boldsymbol{\theta}}(\mathbf{Y}_i) - E^*\{f(X_i, \boldsymbol{\theta})|\mathbf{Y}_i\}$.

4. Solve (12) to obtain $\widehat{\boldsymbol{\theta}}$.

Various proposals for η^* lead to different consistent estimators. Within this class of estimators, the optimal one occurs when $\eta^*(\cdot) = \eta_0(\cdot)$. Hence, a practical and reasonable approach is to propose an $\eta^*(\cdot)$ based on some averaged observations $\widetilde{Y}_{i,\cdot} = \sum_{j=1}^{m} \omega_j Y_{i,j}$. The optimal weights ω_j's naturally should minimize the variance of $\widetilde{Y}_{i,\cdot}$, and it can be easily verified to be $\omega_j = a_j^{-2}/(\sum_{j=1}^{m} a_j^{-2})$.

5.2 Implementation

In implementing the algorithm, the integral equation (11) can be solved using various numerical methods, for example, discretization, to convert it to a problem of solving a linear system. The computation of $E^*\{f(X, \boldsymbol{\theta})|\mathbf{Y}\}$ can be typically performed by the approximation

$$E^*\{f(X, \boldsymbol{\theta})|\mathbf{Y}\} \approx \frac{\sum_{\ell=1}^{K} f(s_\ell, \boldsymbol{\theta}) p(\mathbf{Y}|s_\ell) w_\ell}{\sum_{\ell=1}^{K} p(\mathbf{Y}|s_\ell) w_\ell},$$

where the s_ℓ's are the support points for X and the w_ℓ's are weights we choose to approximate the proposed $\eta^*(X)$, and K is the total number of approximation points that we take.

The computation of conditional expectations $E(\cdot|X)$ is more challenging, especially when m is large, because it involves an m-dimensional integration. Although many computational methods exist to compute multiple dimensional integration, the nature of the problem itself dictates that they are all highly time consuming. Incorporation of such an integration procedure in an estimating equation solving procedure demands even more computational capacity. Thus, direct calculation of (11) is not really feasible.

To lower the dimensionality, we propose to separate the m components of $\mathbf{Y}_i$ into either two or three groups, and use the average value of each group as if they formed the observed data. If the original $Y_{i,j}$ has variance $g(X_i, \theta)$, then the average observation in the k^{th} group has variance $a_k^2 g(X_i, \theta)$, where a_k^2 is the inverse of the number of observations in the k^{th} group. This is why we considered the problem in a more general form than (1) in this section. With this convention in mind, solving (11) is relatively straightforward. We then have the following result.

Theorem 3. *Under regularity conditions, the estimator in (12) is asymptotically consistent, and*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to N(0, A^{-1}BA^{-T})$$

in distribution when $n \to \infty$, where $A = -E(\partial S^*_{eff}/\partial \boldsymbol{\theta})$ and $B = E(S^*_{eff} S^{*T}_{eff})$. If $\eta^* = \eta_0$, then $A = B$ and $A^{-1}BA^{-T} = B^{-1}$ achieves the optimal semiparametric efficiency bound.

Here, the regularity conditions mainly include some sufficient smoothness conditions to permit differentiation and the exchange of differentiation and expectation. It also includes some nonsingularity conditions of the variance matrix to exclude the existence of superefficient estimators. Details of these regularity conditions can be found in Newey (1990). The variance matrix $A^{-1}BA^{-T}$ can be estimated via sample average to compute A and B. In choosing η^* in our problem, we could use $\overline{Y}_{i,\cdot}, i = 1, \ldots, n$ to obtain an approximation of η_0 and proceed with the estimator. Although the approximation is not a valid consistent estimator of η_0, it usually provides a reasonable approximation. The proof of Theorem 3 is in the Supplemental Materials.

6 Application

We applied our methods to the leukemia data from high-density Affymetrix oligonucleotide arrays (Golub et al., 1999). After preprocessing and filtering as in Golub et al. (1999), the data consists of expression level of 3051 genes from 38 bone marrow samples: 27 ALL (acute lymphocytic leukemia) and 11 AML (acute myelogenous leukemia). The data were calibrated and background corrected. To remove possible artifacts due to arrays, as in Huang and Pan (2002), observations on each array are standardized by subtracting the median expression level and dividing by the inter-quantile range of the expression levels on that array. To avoid negative values in the expression level, we then subtract the smallest value across all tumors and all genes from the data set.

Strimmer (2003) used this data to illustrate a quasi-likelihood approach to the estimation of parameters in the quadratic model. To illustrate our methods, we select two subsets, one with two

tumor samples (tumors 1 and 27) and the other with five tumor samples (tumors 1, 8, 13, 21 and 27), from 27 ALL samples. We fit the quadratic model to these two subsets using the naive moment, SIMEX moment and semiparametric methods, where the η^* function used in the semiparametric method is the result of a nonparametric estimation of the averaged tumor sample densities. Observations and fits are shown in Figure 1. As expected, the naive estimator underestimates the trend. Estimates based on other subsets of ALL samples behave similarly.

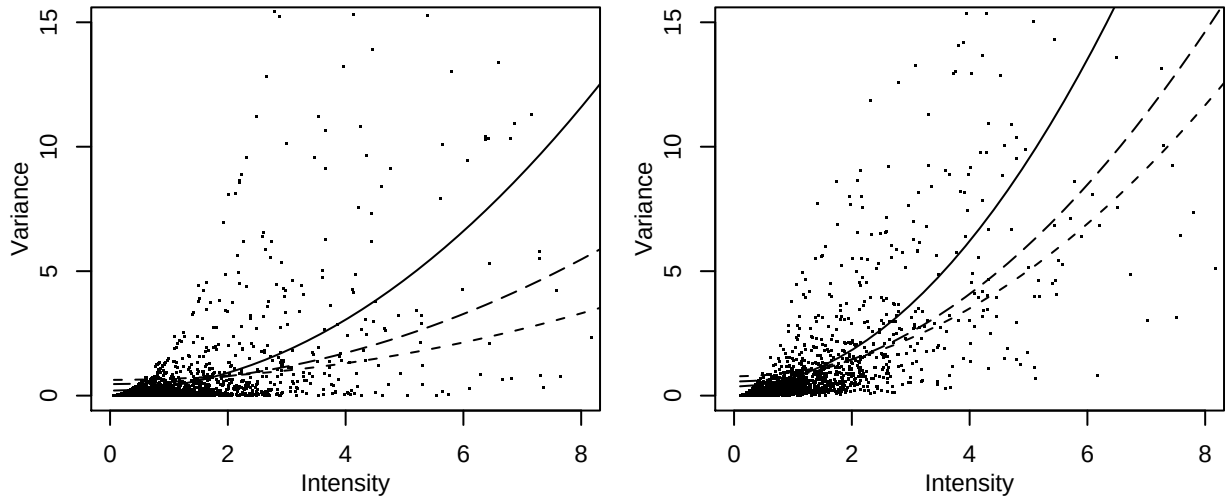

Figure 1: Leukemia data illustration. Left: plot of the subset with two tumor samples. Right: plot of the subset with five tumor samples. Points: sample variances vs sample means. Dashed, long dashed and solid lines are the naive moment, SIMEX moment and semiparametric estimates, respectively.

R, Matlab and Fortran codes have been developed for computing SIMEX and semiparametric estimators. These codes are available from the first author. The CPU times required for computing the parameter estimators are reasonable. For example, permutation SIMEX estimator for 5 selected tumors took about 17 seconds CPU time. The semiparametric method took 7 and 98 seconds respectively for 2 and 3 group estimators.

7 Simulations

We conducted two simulation experiments, one for the constant coefficient of variation model and one for the quadratic variance-mean model. For all simulations, we set $B = 200$ for SIMEX methods.

7.1 Simulations in the Constant Coefficient of Variation Model

For the constant coefficient of variation model, we generate 100 simulation data sets from model (2) with $X_i = \text{Uniform}[1, 3]$. We used a factorial design with $n = 250$ or 500, $m = 3$ or 9, and $\theta = .25$ or 1.

Tables 1 and 2 list squared biases, empirical variances, average of the estimated variances and empirical mean squared errors. Labels "N.M" and "N.R" correspond to the naive moment and regression estimators, "C.M" and "C.R" correspond to the corrected moment and regression estimators defined in (A.3) and (A.4), "S.M" and "S.R" correspond to the SIMEX moment and regression estimators, "PS.M" and "PS.R" correspond to the permutation SIMEX moment and regression estimators, and "Semi.3" and "Semi.2" correspond to the semiparametric estimator, where $\mathbf{Y}_i$ is partitioned to three and two groups respectively. To emphasize the semiparametric estimator's ability to tolerate a misspecified model η^*, we used a normal model with a pre-fixed mean 2 and variance 1 in the simulation.

The simulation confirms the asymptotic results: (a) the naive approach leads to underestimation for both moment and regression estimators; (b) the SIMEX approach corrects bias in the moment estimator, but does not corrects bias in the regression estimator and is badly biased when $m = 3$; (c) the permutation SIMEX and semiparametric approaches corrects bias in both the moment and regression estimators. Bias in the SIMEX regression estimator increases with θ and decreases with m.

It is clear from Tables 1 and 2, that the moments-based approaches are all considerably more efficient that the regression-based approaches. The SIMEX-type moments-based approaches are sometimes less efficient and sometimes more efficient than the semiparametric approaches. It is striking that the moments-based permutation SIMEX method is so competitive with the semiparametric method. When $\mathbf{Y}_i$'s are partitioned into three groups, the semiparametric estimation improved over using two groups. However, such improvement requires much more computation, hence in practice, one may be content with a two group partition.

For the SIMEX and permutation SIMEX, we used the bootstrap procedure to estimate variances of the moment and regression estimators. Variances of the semiparametric estimators were estimated directly. Specifically, consider a data matrix with elements $Y_{i,j}$ given in model (1). The $1,000$ bootstrap samples were generated by resampling rows of the data matrix with replacement. The SIMEX/permutation SIMEX moment/regression estimates were then computed for each boot-

	BSQ	VAR	EVAR	MSE	BSQ	VAR	EVAR	MSE
	$m = 3$				$m = 9$			
	$\theta = .25$, $n = 250$							
N.M	2.84	3.63	–	6.47	0.42	0.92	–	1.34
N.R	47.45	4.21	–	51.66	9.18	1.33	–	10.51
C.M	0.09	5.05	–	5.13	0.00	1.02	–	1.02
C.R	0.60	17.78	–	18.38	0.00	2.24	–	2.24
S.M	0.03	4.91	5.00	4.95	0.00	1.01	1.24	1.01
S.R	9.18	9.27	8.83	18.45	0.19	2.09	2.23	2.28
PS.M	0.10	5.08	5.14	5.18	0.00	1.02	1.24	1.03
PS.R	0.32	12.00	11.90	12.32	0.00	2.28	2.35	2.28
Semi.3	0.00	3.06	3.09	3.07	0.00	2.91	2.89	2.91
Semi.2	0.01	4.89	5.02	4.90	0.01	5.63	5.43	5.64
	$\theta = .25$, $n = 500$							
N.M	3.58	1.68	–	5.25	0.41	0.58	–	0.99
N.R	49.71	1.33	–	51.04	8.82	0.75	–	9.57
C.M	0.00	2.31	–	2.31	0.00	0.64	–	0.64
C.R	0.11	5.04	–	5.15	0.00	1.26	–	1.26
S.M	0.00	2.27	2.39	2.27	0.00	0.64	0.60	0.64
S.R	10.59	2.72	4.22	13.30	0.12	1.19	1.09	1.31
PS.M	0.00	2.32	2.45	2.32	0.00	0.64	0.60	0.64
PS.R	0.04	4.15	5.62	4.18	0.00	1.20	1.14	1.20
Semi.3	0.00	1.59	1.55	1.59	0.00	1.53	1.45	1.53
Semi.2	0.00	2.72	2.52	2.72	0.00	3.05	2.74	3.05

Table 1: Squared bias (BSQ), sample variance (VAR), estimated variance (EVAR) and mean squared error (MSE) for estimates of θ in the constant coefficient of variation model (2) as described in Section 7.1 when $\theta = .25$. All quantities are multiplied by 10000. The symbols "N", "C", "S" and "PS" refer to the naive, corrected, SIMEX and Permutation-SIMEX estimators, respectively. "M" and "R" refer to moments and least square estimators. Semi.j is the semiparametric estimator partitioned into j groups.

	BSQ	VAR	EVAR	MSE	BSQ	VAR	EVAR	MSE
	$m = 3$				$m = 9$			
	$\theta = 1$, $n = 250$							
N.M	6.24	0.45	–	6.69	0.78	0.27	–	1.05
N.R	36.07	0.38	–	36.45	11.49	0.26	–	11.76
C.M	0.00	1.44	–	1.44	0.02	0.41	–	0.43
C.R	0.34	18.56	–	18.90	0.06	1.39	–	1.45
S.M	0.30	1.04	1.27	1.34	0.01	0.40	0.40	0.41
S.R	23.64	1.32	1.18	24.96	1.79	0.69	0.67	2.49
PS.M	0.00	1.43	1.88	1.44	0.03	0.42	0.42	0.44
PS.R	0.00	4.33	5.28	4.33	0.06	1.13	1.19	1.20
Semi.3	0.00	0.83	0.89	0.83	0.00	0.52	0.53	0.52
Semi.2	0.00	1.06	1.09	1.06	0.00	0.77	0.81	0.77
	$\theta = 1$, $n = 500$							
N.M	6.30	0.21	–	6.51	0.99	0.16	–	1.15
N.R	35.14	0.20	–	35.34	11.85	0.15	–	12.01
C.M	0.00	0.67	–	0.67	0.00	0.25	–	0.25
C.R	0.51	7.36	–	7.88	0.01	0.84	–	0.85
S.M	0.32	0.47	0.62	0.79	0.00	0.24	0.19	0.24
S.R	22.26	0.68	0.59	22.94	1.96	0.40	0.33	2.36
PS.M	0.00	0.68	0.89	0.68	0.00	0.25	0.20	0.25
PS.R	0.03	2.13	2.44	2.16	0.02	0.69	0.56	0.71
Semi.3	0.00	0.42	0.44	0.42	0.00	0.27	0.26	0.27
Semi.2	0.00	0.58	0.54	0.58	0.00	0.44	0.41	0.44

Table 2: Squared bias (BSQ), sample variance (VAR), estimated variance (EVAR) and mean squared error (MSE) for estimates of θ in the constant coefficient of variation model (2) as described in Section 7.1 when $\theta = 1$. All quantities are multiplied by 100. The symbols "N", "C", "S" and "PS" refer to the naive, corrected, SIMEX and Permutation-SIMEX estimators, respectively. "M" and "R" refer to moments and least square estimators. Semi.j is the semiparametric estimator partitioned into j groups.

strap sample. The variances of these estimates were used as the bootstrap estimates of variances. The average of the estimated variances are also listed in Tables 1 and 2. The estimated variances all match reasonably well with the empirical variances.

7.2 Simulations in the Quadratic Variation Model

In a simulation not reported here, where the true error distribution is normal, we have observed that the permutation SIMEX method and the semiparametric method had similar performance.

However, to conduct a simulation that is based upon an actual data application, we use the leukemia data in Section 6 to create simulation settings. We first created $\mathcal{X}$ as the collection of gene-specific sample means from the 27 ALL samples with a few very large values excluded. We then fit a quadratic variance model to the 27 ALL samples and create $\mathcal{R}$ as centered and scaled residuals. We note that the distribution of $\mathcal{R}$ is asymmetric and has a heavy right tail. Therefore, the assumptions made in our theory do not hold and this simulation provides a challenge to our methods.

We generated 1000 simulation data sets according to model (1) with X_i sampled with replacement from $\mathcal{X}$, $g(x) = .2037 + .1779x^2$ which is the semiparametric estimate based on tumors 1 and 27 (solid line in the left panel of Figure 1), and $\epsilon_{i,j}$ sampled with replacement from $\mathcal{R}$. We set $n = 3051$, the sample size of the leukemia data, and $m = 5, 10, 15, 20$.

Note that equation (8) is derived based on the normality assumption and especially on fourth moments. As expected, we find that moment estimators have large biases in this simulation since the distribution of $\mathcal{R}$ is far from normal. Therefore, the moment estimator is excluded. The asymptotic extrapolant functions for $\widehat{\alpha}_{PS.R}$ and $\widehat{\beta}_{PS.R}$ have the non-linear form $(a+b\zeta+c\zeta^2)/(c+d\zeta+\zeta^2)$ which is difficult to implement. We tested various lower order polynomial approximations and found that quartic polynomial functions have the best overall performance for $\widehat{\alpha}_{PS.R}$ and $\widehat{\beta}_{PS.R}$. Therefore, in our simulations, quartic functions are used for permutation SIMEX regression estimators. To implement the semiparametric method, we used a gamma model for $\eta^*(X)$ with mean and variance estimated from the generated samples. Such a gamma model allows for the heavier right tail we see in $\mathcal{R}$.

Table 3 lists squared biases, variances and the mean squared errors. Both naive and ordinary SIMEX estimators overestimate α and underestimate β, drastically so when m is small. The bias is especially large for the SIMEX estimator. Both findings are consistent with the simulation in

Section 7.1. The permutation SIMEX estimator reduces both biases and variances in the estimates of β, and the reduction is still substantial even when m equals 20. The permutation SIMEX estimator increases the biases in the estimates of α, especially when m is small. Nevertheless, except for $m = 5$, the permutation SIMEX estimator reduces variances and thus the MSEs of the estimates of α.

The semiparametric estimator provides excellent estimation in terms of mean squared errors for all m values. Noting that the table gives squared bias multiplied by 100, we see that even for $m = 20$ it has an approximate 13% bias for estimating β. The gain in MSE is largely through greatly decreased variability.

It is interesting to see what one would get from the regression-type approaches that SIMEX is based on. To do this, we also computed least squares estimates in the ideal situation where the true means X's are used as the regressor. This ideal estimator serves as a benchmark. It has small biases but variances that are approximately the same as that of permutation SIMEX. Indeed, the semiparametric estimator performs better than the ideal regression estimator with X known for all m. This is not surprising since the least squares used to derive all estimators except the semiparametric estimator are sensitive to very large values in $\mathcal{R}$. The performance of the permutation SIMEX estimator is close to that of the ideal estimator when $m = 15$ or 20. Therefore, it is likely that the performance of the permutation SIMEX estimator is caused by non-robustness associated with the least square method and the fact that $\mathcal{R}$ has a heavy right tail. A robust approach could have been used, although this is beyond the scope of this article. A comparison with the ideal estimator indicates that the performance of both permutation SIMEX and the semiparametric methods are acceptable, even when the distributional assumption about the random errors is violated. We would like to caution that such seemingly robust behavior does not have a theoretical justification and further research is needed before similar behavior can be expected in general situations.

7.3 Simulation Conclusions and Recommendations

There are a few major points that can be gleaned from these simulations. The bias in the naive moment and LS estimators can be substantial, whether the error distribution is normal or not. When the errors are normal, both permutation SIMEX and the semiparametric methods perform well. Which one works better depends on factors such as m. In practice, the errors could be far from normal. In this case, the permutation SIMEX method performs well for the least squares

	BSQ	VAR	MSE	BSQ	VAR	MSE
	$\alpha = 0.2037$			$\beta = 0.1779$		
	$m = 5$					
N.R	1.3433	4.4321	5.7754	12.4622	23.1686	35.6308
S.R	65.0631	214.7426	279.8057	11.0667	35.6630	46.7297
PS.R	6.3592	19.3418	25.7011	0.9517	3.4602	4.4119
Semi.2	0.2218	0.8070	1.0288	0.0286	2.2274	2.2559
I.R	0.0037	1.7305	1.7342	0.0023	0.7512	0.7536
	$m = 10$					
N.R	0.1874	0.9545	1.1419	2.4045	3.9017	6.3062
S.R	6.1626	9.9917	16.1543	1.0934	2.2355	3.3289
PS.R	0.3231	0.4425	0.7656	0.0555	0.1278	0.1833
Semi.2	0.1008	0.2546	0.3554	0.0759	0.0630	0.1390
I.R	0.0001	0.4850	0.4851	0.0000	0.2314	0.2314
	$m = 15$					
N.R	0.0379	0.6179	0.6558	0.8797	2.3901	3.2698
S.R	2.2256	4.7006	6.9262	0.3842	1.1034	1.4876
PS.R	0.0404	0.4312	0.4716	0.0088	0.1575	0.1663
Semi.2	0.0053	0.1214	0.1268	0.0429	0.0457	0.0886
I.R	0.0004	0.3912	0.3916	0.0003	0.1823	0.1827
	$m = 20$					
N.R	0.0081	0.3162	0.3243	0.4198	1.0678	1.4876
S.R	1.1577	1.5012	2.6589	0.2022	0.4225	0.6247
PS.R	0.0101	0.2219	0.2320	0.0021	0.1019	0.1040
Semi.2	0.0020	0.1017	0.1037	0.0515	0.0365	0.0880
I.R	0.0000	0.2659	0.2659	0.0000	0.1273	0.1273

Table 3: Squared bias (BSQ), variance (VAR), and mean squared error (MSE) for the data-based simulation of a quadratic variance function described in Section 7.2. All quantities are multiplied by 100. Sample size is $n = 3051$. The symbols "N", "S", "PS" and "I" refer to the naive, SIMEX, Permutation-SIMEX and Ideal estimators, respectively, where Ideal here means that X is observed. "R" refer to the least square estimators. Semi.2 is the semiparametric estimator partitioned into 2 groups.

estimator, however a robust estimator should be used. Whether the SIMEX and permutation SIMEX methods will reduce bias in a robust estimator remain to be investigated, both in theory and simulations. The semiparametric method performs well in the real data simulation, at least when the chosen distribution for the errors is tuned to the data. It has the ability to be adapted to other specified error distributions. For now, the semiparametric method is recommended. We note that both permutation-SIMEX and the semiparametric method can be improved further: a better extrapolant can improve the performance of permutation SIMEX, and a better approximation of the high dimensional integrals can improve the performance of the semiparametric method.

8 Discussion

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 and Daudin 2005) and data-driven Haar-Fisz (Motakis, Nason, Fryzlewicz and Rutter 2006).

Acknowledgments

Wang's research was supported by a grant from the National Science Foundation (DMS-0706886). Ma's research was supported by a National Science Foundation of Switzerland. Carroll's research was supported by a grant from the National Cancer Institute (CA-57030, CA104620). Carroll's research was supported by grants from the National Cancer Institute (CA57030, CA104620). Part of the work was based on work supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

We thank Dr. Strimmer for sending us the leukemia data. We also thank the editor, associate editor and two referees for constructive comments that substantially improved an earlier draft.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, The Johns Hopkins University Press, Baltimore.

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

Carroll, R. J., Lombard, F., Kuechenhoff, H. and Stefanski, L. A. (1996). Asymptotics for the simex estimator in structural measurement error models, *Journal of the American Statistical Association* **91**: 242–250.

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

Chen, Y., Dougherty, E. R. and 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. and 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. and Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates, *Biostatistics* **6**: 59–75.

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

Devanarayan, V. and Stefanski, L. A. (2002). Empirical simulation extrapolation for measurement error models with replicate measurements, *Statistics and Probability Letters* **59**: 219–225.

Huang, X. and 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. and 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. and 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. and Cavalieri, D. (2003). Fundamentals of cDNA microarray data analysis, *TRENDS in Genetics* **11**: 649–659.

Lin, Y., Nadler, S. T., Attie, A. D. and 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. and 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. and Rutter, G. A. (2006). Variance stabilization and normalization for one-color microarray data using a data-driven multiscale approach, *Bioinformatics* **22**: 2547–2553.

Newey, W. K. (1990). Semiparametric efficiency bounds, *Journal of Applied Econometrics* **5**: 99–135.

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

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

Storey, J. and 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*.

Tong, T. and 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. (2006). *Semiparametric Theory and Missing Data*, Springer, New York.

Tsiatis, A. A. and 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. and Bassett, D. E. (2006). Rosetta error model for gene expression analysis, *Bioinformatics* pp. 1111–1121.

Appendix: Theoretical Results

This Appendix states the major results, with some derivations. Derivations that are largely algebraic in nature are included in Supplemental Materials available at http://www.pstat.ucsb.edu/faculty/yuedong. The main techniques used in the proofs are law of large numbers and the central limit theorem.

A.1 Limiting Results for the Constant Coefficient of Variation Model (2)

Lemma 1. The naive approach that replaces X_i by $\overline{Y}_{i,\cdot}$ has moments and regression estimates that have limiting values

$$\widehat{\theta}_{N.M} \triangleq \frac{n^{-1}\sum_{i=1}^n S_i}{n^{-1}\sum_{i=1}^n \overline{Y}_{i,\cdot}^2} \xrightarrow{p} \frac{\theta}{1+\theta/m}; \tag{A.1}$$

$$\widehat{\theta}_{N.R} \triangleq \frac{n^{-1}\sum_{i=1}^n \overline{Y}_{i,\cdot}^2 S_i}{n^{-1}\sum_{i=1}^n \overline{Y}_{i,\cdot}^4} \xrightarrow{p} \theta\frac{1+\theta/m}{1+6\theta/m+3\theta^2/m^2}. \tag{A.2}$$

Based on (A.1) and (A.2), simple corrections to the naive moment and regression estimators are (taking positive solution in the regression estimator)

$$\widehat{\theta}_{C.M} = \frac{\widehat{\theta}_{N.M}}{1-\widehat{\theta}_{N.M}/m}; \tag{A.3}$$

$$\widehat{\theta}_{C.R} = \frac{-(m-6\widehat{\theta}_{N.R})+\sqrt{m^2-8m\widehat{\theta}_{N.R}+24\widehat{\theta}_{N.R}^2}}{2(1-3\widehat{\theta}_{N.R}/m)}. \tag{A.4}$$

It is easy to see that $\widehat{\theta}_{C.M}$ is consistent. These corrected estimators work very well in simulations (see Tables 1 and 2). However, it is difficult to get simple corrections for other estimators.

Lemma 2. The SIMEX approach has moments and regression estimates that have limiting values

$$\begin{aligned}
\widehat{\theta}_{S.M}(\zeta) &\triangleq B^{-1}\sum_{b=1}^{B}\frac{n^{-1}\sum_{i=1}^{n}S_i}{n^{-1}\sum_{i=1}^{n}W_{b,i}^2(\zeta)} \xrightarrow{p} \frac{\theta}{1+(1+\zeta)\theta/m};\\
\widehat{\theta}_{S.R}(\zeta) &\triangleq B^{-1}\sum_{b=1}^{B}\frac{n^{-1}\sum_{i=1}^{n}W_{b,i}^2(\zeta)S_i}{n^{-1}\sum_{i=1}^{n}W_{b,i}^4(\zeta)} \xrightarrow{p} \theta\frac{1+\theta\left\{(1+\zeta)/m+2\zeta/m(m-1)\right\}}{1+6\theta(1+\zeta)/m+3\theta^2(1+\zeta)^2/m^2}.
\end{aligned}$$

Lemma 3. The permutation SIMEX approach has moments and regression estimates that have limiting values

$$\widehat{\theta}_{PS.M}(\zeta) \triangleq B^{-1}\sum_{b=1}^{B}m^{-1}\sum_{j=1}^{m}\frac{n^{-1}\sum_{i=1}^{n}S_i^{(j)}}{n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^2} \xrightarrow{p} \theta\frac{1+(1+\zeta)/(m-1)}{1+(1+\zeta)\theta/(m-1)}; \tag{A.5}$$

$$\begin{aligned}
\widehat{\theta}_{PS.R}(\zeta) &\triangleq B^{-1}\sum_{b=1}^{B}m^{-1}\sum_{j=1}^{m}\frac{n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^2S_i^{(j)}}{n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^4}\\
&\xrightarrow{p} \theta\frac{1+(1+\theta)(1+\zeta)/(m-1)+3\theta(1+\zeta)^2/(m-1)^2}{1+6\theta(1+\zeta)/(m-1)+3\theta^2(1+\zeta)^2/(m-1)^2}.
\end{aligned} \tag{A.6}$$

A.2 Limiting Results for the Quadratic Variation Model (3)

Lemma 4. The naive approach that replaces X_i by $\overline{Y}_{i,\cdot}$ has moments and regression estimates that have limiting values

$$\begin{aligned}
\widehat{\alpha}_{N.M} &\triangleq n^{-1}\sum_{i=1}^{n}S_i-\widehat{\beta}_{N.M}n^{-1}\sum_{i=1}^{n}\overline{Y}_{i,\cdot}^2 \xrightarrow{p} \alpha+A;\\
\widehat{\beta}_{N.M} &\triangleq \sqrt{\frac{\frac{m-1}{m+1}n^{-1}\sum_{i=1}^{n}S_i^2-(n^{-1}\sum_{i=1}^{n}S_i)^2}{n^{-1}\sum_{i=1}^{n}\overline{Y}_{i,\cdot}^4-(n^{-1}\sum_{i=1}^{n}\overline{Y}_{i,\cdot}^2)^2}} \xrightarrow{p} \beta\sqrt{\frac{\mathrm{var}(X^2)}{\mathrm{var}(X^2)+C}};\\
\widehat{\alpha}_{N.R} &\triangleq n^{-1}\sum_{i=1}^{n}S_i-\widehat{\beta}_{N.R}n^{-1}\sum_{i=1}^{n}\overline{Y}_{i,\cdot}^2 \xrightarrow{p} \alpha+D;\\
\widehat{\beta}_{N.R} &\triangleq \frac{n^{-1}\sum_{i=1}^{n}\overline{Y}_{i,\cdot}^2S_i-(n^{-1}\sum_{i=1}^{n}\overline{Y}_{i,\cdot}^2)(n^{-1}\sum_{i=1}^{n}S_i)}{n^{-1}\sum_{i=1}^{n}\overline{Y}_{i,\cdot}^4-(n^{-1}\sum_{i=1}^{n}\overline{Y}_{i,\cdot}^2)^2} \xrightarrow{p} \beta\frac{(1+\beta/m)\mathrm{var}(X^2)}{\mathrm{var}(X^2)+C},
\end{aligned}$$

where

$$\begin{aligned}
A &= \beta\mathrm{E}(X^2)\left\{1-\sqrt{\frac{\mathrm{var}(X^2)}{\mathrm{var}(X^2)+C}}\right\}-\frac{\beta}{m}\mathrm{E}\{g(X)\}\sqrt{\frac{\mathrm{var}(X^2)}{\mathrm{var}(X^2)+C}},\\
C &= m^{-1}\left\{4\alpha\mathrm{E}(X^2)+4\beta\mathrm{E}(X^4)+2\beta\mathrm{var}(X^2)+\frac{2\alpha^2}{m}+\frac{4\alpha\beta}{m}\mathrm{E}(X^2)\right.\\
&\quad\left.+\frac{\beta^2}{m}\mathrm{var}(X^2)+\frac{2\beta^2}{m}\mathrm{E}(X^4)\right\},\\
D &= \frac{\beta(mC-\beta\mathrm{var}\{X^2\})\mathrm{E}(X^2)-\beta(1+\beta/m)\{\alpha+\beta\mathrm{E}(X^2)\}\mathrm{var}(X^2)}{m\{\mathrm{var}(X^2)+C\}}.
\end{aligned}$$

Lemma 5. The SIMEX approach has moments and regression estimates that have limiting values

$$\begin{aligned}
\widehat{\alpha}_{S.M}(\zeta) &\triangleq n^{-1}\sum_{i=1}^{n} S_i - B^{-1}\sum_{b=1}^{B}\widehat{\beta}_{S.M}(\zeta)n^{-1}\sum_{i=1}^{n} W_{b,i}^2(\zeta) \xrightarrow{p} \alpha + F;\\
\widehat{\beta}_{S.M}(\zeta) &\triangleq B^{-1}\sum_{b=1}^{B}\sqrt{\frac{\frac{m-1}{m+1}n^{-1}\sum_{i=1}^{n} S_i^2 - (n^{-1}\sum_{i=1}^{n} S_i)^2}{n^{-1}\sum_{i=1}^{n} W_{b,i}^4(\zeta) - \{n^{-1}\sum_{i=1}^{n} W_{b,i}^2(\zeta)\}^2}} \xrightarrow{p} \beta\sqrt{\frac{\mathrm{var}(X^2)}{\mathrm{var}(X^2)+(1+\zeta)G}};\\
\widehat{\alpha}_{S.R}(\zeta) &\triangleq n^{-1}\sum_{i=1}^{n} S_i - B^{-1}\sum_{b=1}^{B}\widehat{\beta}_{S.R}(\zeta)n^{-1}\sum_{i=1}^{n} W_{b,i}^2(\zeta) \xrightarrow{p} \alpha + H;\\
\widehat{\beta}_{S.R}(\zeta) &\triangleq B^{-1}\sum_{b=1}^{B}\frac{n^{-1}\sum_{i=1}^{n} W_{b,i}^2(\zeta)S_i - \{n^{-1}\sum_{i=1}^{n} W_{b,i}^2(\zeta)\}(n^{-1}\sum_{i=1}^{n} S_i)}{n^{-1}\sum_{i=1}^{n} W_{b,i}^4(\zeta) - \{n^{-1}\sum_{i=1}^{n} W_{b,i}^2(\zeta)\}^2}\\
&\xrightarrow{p} \frac{\beta\mathrm{var}(X^2) + \beta^2(1+\zeta)\mathrm{var}(X^2)/m + 2\zeta\mathrm{E}\{g^2(X)\}/\{m(m-1)\}}{\mathrm{var}(X^2)+(1+\zeta)G},
\end{aligned}$$

where

$$\begin{aligned}
F &= \beta\mathrm{E}(X^2)\left\{1 - \sqrt{\frac{\mathrm{var}(X^2)}{\mathrm{var}(X^2)+(1+\zeta)G}}\right\} - \frac{\beta(1+\zeta)}{m}\mathrm{E}\{g(X)\}\sqrt{\frac{\mathrm{var}(X^2)}{\mathrm{var}(X^2)+(1+\zeta)G}},\\
G &= \frac{6}{m}\mathrm{E}\{X^2 g(X)\} - \frac{2}{m}\mathrm{E}(X^2)\mathrm{E}\{g(X)\} + \frac{3(1+\zeta)}{m^2}\mathrm{E}\{g^2(X)\} - \frac{(1+\zeta)}{m^2}[\mathrm{E}\{g(X)\}]^2,\\
H &= \beta\mathrm{E}(X^2) - \frac{\beta\mathrm{var}(X^2) + \beta^2(1+\zeta)\mathrm{var}(X^2)/m + 2\zeta\mathrm{E}\{g^2(X)\}/\{m(m-1)\}}{\mathrm{var}(X^2)+(1+\zeta)G}\\
&\quad\times\left[\mathrm{E}(X^2) + \frac{1+\zeta}{m}\mathrm{E}\{g(X)\}\right].
\end{aligned}$$

Lemma 6. The permutation SIMEX approach has moments and regression estimates that have limiting values

$$\begin{aligned}
\widehat{\alpha}_{PS.M}(\zeta) &\triangleq B^{-1}\sum_{b=1}^{B} m^{-1}\sum_{j=1}^{m}\left\{n^{-1}\sum_{i=1}^{n} S_i^{(j)} - \widehat{\beta}_{PS.M}(\zeta)n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^2\right\} \xrightarrow{p} \alpha + (1+\zeta)I;\\
\widehat{\beta}_{PS.M}(\zeta) &\triangleq B^{-1}\sum_{b=1}^{B} m^{-1}\sum_{j=1}^{m}\sqrt{\frac{\frac{1}{3n}\sum_{i=1}^{n}(S_i^{(j)})^2 - (n^{-1}\sum_{i=1}^{n} S_i^{(j)})^2}{n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^4 - [n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^2]^2}}\\
&\xrightarrow{p} \beta\sqrt{\frac{\mathrm{var}(X^2)+(1+\zeta)J}{\mathrm{var}(X^2)+(1+\zeta)K}};\\
\widehat{\alpha}_{PS.R}(\zeta) &\triangleq B^{-1}\sum_{b=1}^{B} m^{-1}\sum_{j=1}^{m}\left[n^{-1}\sum_{i=1}^{n} S_i^{(j)} - \widehat{\beta}_{PS.R}n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^2\right] \xrightarrow{p} \alpha + (1+\zeta)L;\\
\widehat{\beta}_{PS.R}(\zeta) &\triangleq B^{-1}\sum_{b=1}^{B} m^{-1}\sum_{j=1}^{m}\frac{n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^2 S_i^{(j)} - [n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^2](n^{-1}\sum_{i=1}^{n} S_i^{(j)})}{n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^4 - [n^{-1}\sum_{i=1}^{n}\{W_{b,i}^{(j)}(\zeta)\}^2]^2}\\
&\xrightarrow{p} \frac{\beta\mathrm{var}(X^2)+(1+\zeta)M}{\mathrm{var}(X^2)+(1+\zeta)K},
\end{aligned}$$

where

$$
\begin{aligned}
I &= \beta \mathrm{E}(X^2)\left\{1-\sqrt{\frac{\mathrm{var}(X^2)+(1+\zeta)J}{\mathrm{var}(X^2)+(1+\zeta)K}}\right\} \\
&\quad +\frac{1+\zeta}{m-1}\mathrm{E}\{g(X)\}\left\{1-\beta\sqrt{\frac{\mathrm{var}(X^2)+(1+\zeta)J}{\mathrm{var}(X^2)+(1+\zeta)K}}\right\}, \\
J &= \frac{6}{m-1}\mathrm{E}\{g^2(X)\}-\frac{2}{m-1}[\mathrm{E}\{g(X)\}]^2+\frac{3(1+\zeta)}{(m-1)^2}\mathrm{E}\{g^2(X)\} \\
&\quad -\frac{(1+\zeta)}{(m-1)^2}[\mathrm{E}\{g(X)\}]^2, \\
K &= \frac{6}{m-1}\mathrm{E}\{X^2g(X)\}-\frac{2}{m-1}\mathrm{E}(X^2)\mathrm{E}\{g(X)\}+\frac{3(1+\zeta)}{(m-1)^2}\mathrm{E}\{g^2(X)\} \\
&\quad -\frac{(1+\zeta)}{(m-1)^2}[\mathrm{E}\{g(X)\}]^2, \\
L &= \mathrm{E}(X^2)\left\{\beta-\frac{\beta\mathrm{var}(X^2)+(1+\zeta)M}{\mathrm{var}(X^2)+(1+\zeta)K}\right\} \\
&\quad +\frac{1+\zeta}{m-1}\mathrm{E}\{g(X)\}\left\{1-\frac{\beta\mathrm{var}(X^2)+(1+\zeta)M}{\mathrm{var}(X^2)+(1+\zeta)K}\right\}, \\
M &= \frac{\beta+\beta^2}{m-1}\mathrm{var}(X^2)+\frac{(1+\zeta)}{(m-1)^2}\left\{3\mathrm{E}\{g^2(X)\}-[\mathrm{E}\{g(X)\}]^2\right\}.
\end{aligned}
$$

A.3 Asymptotic Normality of SIMEX Estimates

Asymptotic normality for all the estimators follows along the same lines as the asymptotic theory for constant-variation SIMEX (Carroll, Lombard, Kuechenhoff and Stefanski, 1996).

Here we merely sketch the argument for the SIMEX moment estimator $\widehat{\theta}_{S,M}(\zeta)$ for the constant coefficient of variation model (2): all other estimators follow along similar lines. For any fixed b, the estimator $\widehat{\theta}_{S,M,b}(\zeta)$ solves

$$0 = n^{-1}\sum_{i=1}^{n}\{S_i - \widehat{\theta}_{S,M,b}(\zeta)W_{b,i}^2(\zeta)\},$$

and of course $\widehat{\theta}_{S,M}(\zeta) = B^{-1}\sum_{b=1}^{B}\widehat{\theta}_{S,M,b}(\zeta)$. Using Fact 8, $n^{-1}\sum_{i=1}^{n}W_{b,i}^2(\zeta) \stackrel{p}{\to} \{1+(1+\zeta)/m\}E(X^2)$. As seen in Lemma 2,

$$\widehat{\theta}_{S,M,b}(\zeta) \stackrel{p}{\to} \theta_{S.M}(\zeta) = \frac{\theta}{1+(1+\zeta)\theta/m}.$$

Using standard estimating equation calculations, we see that for any fixed b,

$$\{1+(1+\zeta)/m\}E(X^2)n^{1/2}\left\{\widehat{\theta}_{S,M,b}(\zeta)-\theta_{S.M}(\zeta)\right\} = n^{-1/2}\sum_{i=1}^{n}\{S_i-\theta_{S,M}(\zeta)W_{b,i}^2(\zeta)\}+o_p(1).$$

Since B is fixed, this means that

$$n^{1/2}\left\{\widehat{\theta}_{S,M}(\zeta)-\theta_{S.M}(\zeta)\right\} = \frac{n^{-1/2}\sum_{i=1}^{n}\{S_i-\theta_{S,M}(\zeta)B^{-1}\sum_{b=1}^{B}W_{b,i}^2(\zeta)\}}{\{1+(1+\zeta)/m\}E(X^2)}+o_p(1).$$

Since

$$0 = E\{S_i - \theta_{S,M}(\zeta)B^{-1}\sum_{b=1}^{B} W_{b,i}^2(\zeta)\},$$

the central limit theorem shows that for any finite set $(\zeta_1 = 0, ..., \zeta_M)$,

$$\begin{bmatrix} n^{1/2}\left\{\widehat{\theta}_{S,M}(\zeta_1) - \theta_{S.M}(\zeta_1)\right\} \\ \vdots \\ n^{1/2}\left\{\widehat{\theta}_{S,M}(\zeta_M) - \theta_{S.M}(\zeta_M)\right\} \end{bmatrix} \tag{A.7}$$

has a joint multivariate normal limiting distribution. The extrapolated (to $\zeta = -1$) estimators are smooth functions of (A.7), the delta-method shows that the extrapolated estimators are asymptotically normally distributed as well.

It is possible to estimate the joint limiting covariance matrix of (A.7) using the following algorithm. Because from Fact 8 $(nB)^{-1}\sum_{b=1}^{B}\sum_{i=1}^{n} W_{b,i}^2(\zeta)$ is a consistent estimate of $\{1 + (1 + \zeta)/m\}E(X^2)$, a consistent estimate of the asymptotic covariance matrix of (A.7) is just the sample covariance matrix of the terms

$$\begin{bmatrix} \{S_i - \widehat{\theta}_{S,M}(\zeta_1)B^{-1}\sum_{b=1}^{B} W_{b,i}^2(\zeta_1)\}/(nB)^{-1}\sum_{b=1}^{B}\sum_{i=1}^{n} W_{b,i}^2(\zeta_1) \\ \vdots \\ \{S_i - \widehat{\theta}_{S,M}(\zeta_M)B^{-1}\sum_{b=1}^{B} W_{b,i}^2(\zeta_M)\}/(nB)^{-1}\sum_{b=1}^{B}\sum_{i=1}^{n} W_{b,i}^2(\zeta_M) \end{bmatrix}.$$

A.4 Derivation of Λ, Λ^T and S_{eff}

Replacing the nuisance function $\eta(X)$ with $p_X(X, \gamma)$ for some finite dimensional parameter γ, then the score function of $p_{\mathbf{Y}}(\mathbf{Y}, \boldsymbol{\theta}, \gamma)$ with respect to γ has the form

$$S_\gamma = \partial \log \int p_X(X, \gamma) p_{\mathbf{Y}}(\mathbf{Y}, \boldsymbol{\theta}) d\mu(X)/\partial\gamma = E\{\partial \log p_X(X, \gamma)/\partial\gamma | Y\}$$

Since $p_X(X, \gamma)$ is an arbitrary pdf, $\partial \log p_X(X, \gamma)$ can be an arbitrary mean zero function of X. Take into consideration all possible parametrization of $\eta(X)$, we obtain the nuisance tangent space

$$\Lambda = [E\{f(X)|\mathbf{Y}\} : \forall f(X) \text{ s.t. } E(f) = 0].$$

The nuisance tangent space orthogonal complement $\Lambda^\perp$ can be easily verified to be

$$\Lambda^\perp = \{g(\mathbf{Y}) : E(g|X) = 0\}.$$

The projection of the score function to $\Lambda^\perp$, S_{eff} needs to satisfy two conditions, $S_{eff} \in \Lambda^\perp$ and $S_\theta - S_{eff} \in \Lambda$. It can be easily verified that this leads to the expression in (9) and (10).

Supplemental Materials: Sketch of Technical Arguments

The following text is included for refereeing purpose only. It will not be part of the paper.

Some Simple Facts

Assume that X_i are iid random variables which are independent of $\epsilon_{i,j}$, and $\epsilon_{i,j} \overset{iid}{\sim} \mathrm{N}(0,1)$.

1. $\overline{Y}_{i,\cdot} = X_i + g^{1/2}(X_i)\bar{\epsilon}_{i,\cdot}$ where $\bar{\epsilon}_{i,\cdot} = \sum_{j=1}^m \epsilon_{i,j}/m \sim \mathrm{N}(0, 1/m)$.
2. $S_i = g(X_i)S_{i,\epsilon}$ where $S_{i,\epsilon} = \sum_{j=1}^m (\epsilon_{i,j} - \bar{\epsilon}_{i,\cdot})^2/(m-1)$.
3. $n^{-1}\sum_{i=1}^n \overline{Y}_{i,\cdot}^2 \xrightarrow{p} \mathrm{E}(\overline{Y}_{1\cdot}^2) = \mathrm{E}(X_1^2) + \mathrm{E}(g(X_1))/m$.
4. $n^{-1}\sum_{i=1}^n S_i \xrightarrow{p} \mathrm{E}(S_1) = \mathrm{E}(g(X_1))$.
5. $n^{-1}\sum_{i=1}^n \overline{Y}_{i,\cdot}^2 S_i \xrightarrow{p} \mathrm{E}(\overline{Y}_{1\cdot}^2 S_1) = \mathrm{E}(X_1^2 g(X_1)) + \mathrm{E}(g^2(X_1))/m$.
6. $n^{-1}\sum_{i=1}^n \overline{Y}_{i,\cdot}^4 \xrightarrow{p} \mathrm{E}(\overline{Y}_{i,\cdot}^4) = \mathrm{E}(X_1^4) + 6\mathrm{E}(X_1^2 g(X_1))/m + 3\mathrm{E}(g^2(X_1))/m^2$.
7. $W_{b,i}(\zeta) = X_i + g^{1/2}(X_i)\Omega_i$, where $\Omega_i = \bar{\epsilon}_{i,\cdot} + (\zeta/m)^{1/2}\sum_{j=1}^m c_{b,i,j}\epsilon_{i,j} \sim \mathrm{N}(0, (1+\zeta)/m)$.
8. $n^{-1}\sum_{i=1}^n W_{b,i}^2(\zeta) \xrightarrow{p} \mathrm{E}(W_{b,1}^2(\zeta)) = \mathrm{E}(X_1^2) + (1+\zeta)\mathrm{E}(g(X_1))/m$.
9. $n^{-1}\sum_{i=1}^n W_{b,i}^4(\zeta) \xrightarrow{p} \mathrm{E}(W_{b,1}^4(\zeta)) = \mathrm{E}(X_1^4) + 6(1+\zeta)\mathrm{E}(X_1^2 g(X_1))/m + 3(1+\zeta)^2\mathrm{E}(g^2(X_1))/m^2$.
10. $n^{-1}\sum_{i=1}^n W_{b,i}^2(\zeta)S_i \xrightarrow{p} \mathrm{E}(W_{b,1}^2(\zeta)S_1) = \mathrm{E}(X_1^2 g(X_1)) + \{(1+\zeta)/m + 2\zeta/m(m-1)\}\,\mathrm{E}(g^2(X_1))$.
11. $n^{-1}\sum_{i=1}^n S_i^{(1)} \xrightarrow{p} \mathrm{E}(S_1^{(1)}) = (1 + (1+\zeta)/(m-1))\mathrm{E}(g(X_1))$.
12. $n^{-1}\sum_{i=1}^n (W_{b,i}^{(1)}(\zeta))^2 S_i^{(1)} \xrightarrow{p} \mathrm{E}\{(W_{b,1}^{(1)}(\zeta))^2 S_i^{(1)}\}$
    $= \mathrm{E}(X_1^2 g(X_1)) + (1+\zeta)\{\mathrm{E}(X_1^2 g(X_1)) + \mathrm{E}(g^2(X_1))\}/(m-1) + 3(1+\zeta)^2\mathrm{E}(g^2(X_1))/(m-1)^2$.
13. $n^{-1}\sum_{i=1}^n S_i^2 \xrightarrow{p} \mathrm{E}(S_1^2) = (m+1)\mathrm{E}(g^2(X_1))/(m-1)$.
14. $n^{-1}\sum_{i=1}^n (S_i^{(1)})^2 \xrightarrow{p} \mathrm{E}(S_1^{(1)})^2 = 3(1 + (1+\zeta)/(m-1))^2\mathrm{E}(g^2(X_1))$.

[Proof of 10]

$$\mathrm{E}(W_{b,1}^2(\zeta)S_1) = \mathrm{E}(X_1^2 g(X_1)) + 2\mathrm{E}(X_1 g^{3/2}(X_1))\mathrm{E}(\Omega_1 S_{1,\epsilon}) + \mathrm{E}(g^2(X_1))\mathrm{E}(\Omega_1^2 S_{1,\epsilon}).$$

It is easy to see that $\mathrm{E}(\Omega_1 S_{1,\epsilon}) = 0$. Let

$$\Omega_1^2 = \bar{\epsilon}_{1,\cdot}^2 + 2(\frac{\zeta}{m})^{1/2}\bar{\epsilon}_{1,\cdot}\sum_{j=1}^m c_{b,1,j}\epsilon_{1,j} + \frac{\zeta}{m}(\sum_{j=1}^m c_{b,1,j}\epsilon_{1,j})^2 \triangleq A_1 + A_2 + A_3.$$

Since $\bar{\epsilon}_{1,\cdot}$ is independent of $S_{1,\epsilon}$, we have

$$\mathrm{E}(A_1 S_{1,\epsilon}) = m^{-1}. \tag{8}$$

Now

$$\begin{aligned}
\mathrm{E}(A_2 S_{1,\epsilon}) &= \frac{2(\zeta/m)^{1/2}}{m-1}\left\{ m^{-1}\mathrm{E}\left(\sum_{j=1}^{m}\epsilon_{1,j}\sum_{k=1}^{m}c_{b,1,k}\epsilon_{1,k}\sum_{l=1}^{m}\epsilon_{1,l}^2\right) - m\mathrm{E}\left(\bar{\epsilon}_{i,\cdot}^3\sum_{k=1}^{m}c_{b,1,k}\epsilon_{1,k}\right)\right\} \\
&= \frac{2(\zeta/m)^{1/2}}{m(m-1)}\left\{ \mathrm{E}\left(\sum_{j=1}^{m}c_{b,1,j}\epsilon_{1,j}^4\right) + \mathrm{E}\left(\sum_{j=1}^{m}\sum_{l\neq j}c_{b,1,j}\epsilon_{1,j}^2\epsilon_{1,l}^2\right)\right\} \\
&= \frac{2(\zeta/m)^{1/2}}{m(m-1)}\left\{ 3\sum_{j=1}^{m}c_{b,1,j} + (m-1)\sum_{j=1}^{m}c_{b,1,j}\right\} \\
&= 0, \qquad (9)
\end{aligned}$$

where we have used the facts that $\bar{\epsilon}_{1,\cdot}$ and $\sum_{k=1}^{m}c_{b,1,k}\epsilon_{1,k}$ are independent and $\sum_{j=1}^{m}c_{b,1,j}=0$. Finally,

$$\begin{aligned}
\mathrm{E}(A_3 S_{1,\epsilon}) &= \frac{\zeta}{m(m-1)}\mathrm{E}\left\{\left(\sum_{j=1}^{m}c_{b,1,j}\epsilon_{1,j}\right)^2\sum_{j=1}^{m}\epsilon_{1,j}^2 - \left(\sum_{j=1}^{m}c_{b,1,j}\epsilon_{1,j}\right)^2 m\bar{\epsilon}_{1,\cdot}^2\right\} \\
&= \frac{\zeta}{m(m-1)}\left\{\mathrm{E}\left(\sum_{j=1}^{m}c_{b,1,j}^2\epsilon_{1,j}^2\sum_{k=1}^{m}\epsilon_{1,k}^2\right) - m\mathrm{E}\left(\sum_{j=1}^{m}c_{b,1,j}\epsilon_{1,j}\right)^2\mathrm{E}\left(\bar{\epsilon}_{1,\cdot}^2\right)\right\} \\
&= \frac{\zeta}{m(m-1)}\left\{\mathrm{E}\left(\sum_{j=1}^{m}c_{b,1,j}^2\epsilon_{1,j}^4\right) + \mathrm{E}\left(\sum_{j=1}^{m}\sum_{k\neq j}^{m}c_{b,1,j}^2\epsilon_{1,j}^2\epsilon_{1,k}^2\right) - 1\right\} \\
&= \frac{\zeta}{m(m-1)}\left\{3\sum_{j=1}^{m}c_{b,1,j}^2 + (m-1)\sum_{j=1}^{m}c_{b,1,j}^2 - 1\right\} \\
&= \frac{(m+1)\zeta}{m(m-1)}, \qquad (10)
\end{aligned}$$

where we have used the fact that $\sum_{j=1}^{m}c_{b,1,j}^2=1$. Collecting terms, we have the fact 10.

[Proof of 11] Let

$$\Omega_i^{(j)} = \frac{1}{m-1}\sum_{k\neq j}\epsilon_{i,j} + (\frac{\zeta}{m-1})^{1/2}\sum_{k\neq j}c_{b,i,k}^{(j)}\epsilon_{i,k} \sim \mathrm{N}(0, \frac{1+\zeta}{m-1}).$$

$\Omega_i^{(j)}$ is independent of $\epsilon_{i,j}$. Then

$$\begin{aligned}
\mathrm{E}(S_1^{(1)}) &= \mathrm{E}(Y_{1,1} - W_{b,1}^{(1)}(\zeta))^2 \\
&= \mathrm{E}\left\{g(X_1)(\epsilon_{1,1} - \Omega_1^{(1)})^2\right\} \\
&= (1 + \frac{1+\zeta}{m-1})\mathrm{E}(g(X_1)).
\end{aligned}$$

[Proof of 12]

$$\begin{aligned}
& \mathrm{E}\left\{W_{b,1}^{(1)}(\zeta)^2 S_1^{(1)}\right\} \\
= \; & \mathrm{E}\left\{(X_1 + g^{1/2}(X_1)\Omega_1^{(1)})^2(\epsilon_{1,1} - \Omega_1^{(1)})^2 g(X_1)\right\} \\
= \; & \mathrm{E}\left[\left\{X_1^2 + 2X_1 g^{1/2}(X_1)\Omega_1^{(1)} + g(X_1)(\Omega_1^{(1)})^2\right\}\left\{\epsilon_{1,1}^2 - 2\epsilon_{1,1}\Omega_1^{(1)} + (\Omega_1^{(1)})^2\right\} g(X_1)\right] \\
= \; & \mathrm{E}(X_1^2 g(X_1)) + \frac{1+\zeta}{m-1}\mathrm{E}(X_1^2 g(X_1)) + \frac{1+\zeta}{m-1}\mathrm{E}(g^2(X_1)) + \frac{3(1+\zeta)^2}{(m-1)^2}\mathrm{E}(g^2(X_1)) \\
= \; & \mathrm{E}(X_1^2 g(X_1)) + \frac{1+\zeta}{m-1}(\mathrm{E}(X_1^2 g(X_1)) + \mathrm{E}(g^2(X_1))) + \frac{3(1+\zeta)^2}{(m-1)^2}\mathrm{E}(g^2(X_1)).
\end{aligned}$$

[Proof of 14]

$$\begin{aligned}
& \mathrm{E}(Y_{1,1} - W_{b,1}^{(1)}(\zeta))^4 \\
= \; & \mathrm{E}\left\{g^2(X_1)(\epsilon_{1,1} - \Omega_1^{(1)})^4\right\} \\
= \; & \{3 + 6(1+\zeta)/(m-1) + 3(1+\zeta)^2/(m-1)^2\}\mathrm{E}(g^2(X_1)) \\
= \; & 3(1 + (1+\zeta)/(m-1))^2\mathrm{E}(g^2(X_1))
\end{aligned}$$

Proofs of Lemmas 1, 2 and 3

Since $g(x) = \theta x^2$, then $\mathrm{E}\{g(X)\} = \theta\mathrm{E}(X^2)$, $\mathrm{E}\{X^2 g(X)\} = \theta\mathrm{E}(X^4)$ and $\mathrm{E}\{g^2(X)\} = \theta^2\mathrm{E}(X^4)$. Lemma 1 follows from facts 3 to 6. Lemma 2 follows from facts 4 to 8. Lemma 3 follows from facts 8, 9, 11, and 12.

Proof of Lemma 4

Since $g(x) = \alpha + \beta x^2$, then $\mathrm{E}\{g(X)\} = \alpha + \beta\mathrm{E}(X^2)$, $\mathrm{E}\{X^2 g(X)\} = \alpha\mathrm{E}(X^2) + \beta\mathrm{E}(X^4)$, $\mathrm{E}\{g^2(X)\} = \alpha^2 + 2\alpha\beta\mathrm{E}(X^2) + \beta^2\mathrm{E}(X^4)$, $\mathrm{var}\{g(X)\} = \beta^2\mathrm{var}(X^2)$ and $\mathrm{E}\{X^2 g(X)\} - \mathrm{E}(X^2)\mathrm{E}\{g(X)\} = \beta\mathrm{var}(X^2)$.

From facts 4 and 13,

$$\frac{m-1}{m+1} n^{-1}\sum_{i=1}^{n} S_i^2 - (n^{-1}\sum_{i=1}^{n} S_i)^2 \xrightarrow{p} \mathrm{E}(g^2(X)) - (\mathrm{E}(g(X)))^2 = \mathrm{var}(g(X)) = \beta^2\mathrm{var}(X^2). \qquad (11)$$

Using facts 3 and 6,

$$\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\overline{Y}_{i,\cdot}^{4} - (n^{-1}\sum_{i=1}^{n}\overline{Y}_{i,\cdot}^{2})^2 \\
\xrightarrow{p} \;& \mathrm{E}(X^4) + \frac{6}{m}\mathrm{E}(X^2 g(X)) + \frac{3}{m^2}\mathrm{E}(g^2(X)) - \left\{\mathrm{E}(X^2) + m^{-1}\mathrm{E}(g(X))\right\}^2 \\
= \;& \mathrm{var}(X^2) + \frac{4}{m}\mathrm{E}(X^2 g(X)) + \frac{2}{m}\left\{\mathrm{E}(X^2 g(X)) - \mathrm{E}(X^2)\mathrm{E}(g(X))\right\} + \frac{2}{m^2}\mathrm{E}(g^2(X)) \\
& + \frac{1}{m^2}\mathrm{var}(X^2) \\
= \;& \mathrm{var}(X^2) + C. \qquad (12)
\end{aligned}$$

Combining (11) and (12) leads to the result for $\widehat{\beta}_{N.M}$. Now,

$$\begin{aligned}
\widehat{\alpha}_{N.M} \xrightarrow{p} \;& \mathrm{E}(g(X)) - \beta\sqrt{\frac{\mathrm{var}(X^2)}{\mathrm{var}(X^2)+C}}\left\{\mathrm{E}(X^2) + m^{-1}\mathrm{E}(g(X))\right\} \\
= \;& \alpha + \beta\mathrm{E}(X^2)\left\{1 - \sqrt{\frac{\mathrm{var}(X^2)}{\mathrm{var}(X^2)+C}}\right\} - \frac{\beta}{m}\sqrt{\frac{\mathrm{var}(X^2)}{\mathrm{var}(X^2)+C}}\mathrm{E}(g(X)) \\
= \;& \alpha + A.
\end{aligned}$$

Using facts 3 to 5,

$$\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\overline{Y}_{i,\cdot}^{2} S_i - (n^{-1}\sum_{i=1}^{n}\overline{Y}_{i,\cdot}^{2})(n^{-1}\sum_{i=1}^{n} S_i) \\
\xrightarrow{p} \;& \mathrm{E}(X^2 g(X)) + m^{-1}\mathrm{E}(g^2(X)) - \left\{\mathrm{E}(X^2) + m^{-1}\mathrm{E}(g(X))\right\}\mathrm{E}(g(X)) \\
= \;& \mathrm{E}(X^2 g(X)) - \mathrm{E}(X^2)\mathrm{E}(g(X)) + m^{-1}\mathrm{var}(g(X^2)) \\
= \;& \beta\mathrm{var}(X^2) + \frac{\beta^2}{m}\mathrm{var}(X^2) \\
= \;& \beta(1 + \frac{\beta}{m})\mathrm{var}(X^2). \qquad (13)
\end{aligned}$$

Combining (13) and (12) leads to the result for $\widehat{\beta}_{N.R}$. Now,

$$\begin{aligned}
\widehat{\alpha}_{N.R} \xrightarrow{p} \;& \mathrm{E}(g(X)) - \beta\frac{(1+\beta/m)\mathrm{var}(X^2)}{\mathrm{var}(X^2)+C}\left\{\mathrm{E}(X^2) + m^{-1}\mathrm{E}(g(X))\right\} \\
= \;& \alpha + \frac{\beta(mC - \beta\mathrm{var}(X^2))\mathrm{E}(X^2) - \beta(1+\beta/m)\{\alpha + \beta\mathrm{E}(X^2)\}\mathrm{var}(X^2)}{m(\mathrm{var}(X^2)+C)} \\
= \;& \alpha + D.
\end{aligned}$$

Proof of Lemma 5

Using facts 8 and 9,

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n} W_{b,i}^4(\zeta) - (n^{-1}\sum_{i=1}^{n} W_{b,i}^2(\zeta))^2 \\ \xrightarrow{p} & \mathrm{E}(X^4) + \frac{6(1+\zeta)}{m}\mathrm{E}(X^2 g(X)) + \frac{3(1+\zeta)^2}{m^2}\mathrm{E}(g^2(X)) - \left\{\mathrm{E}(X^2) + \frac{1+\zeta}{m}\mathrm{E}(g(X))\right\}^2 \\ = & \mathrm{var}(X^2) + (1+\zeta)G. \end{aligned} \tag{14}$$

Combining (11) and (14) leads to the result for $\widehat{\beta}_{S.M}$. Now

$$\begin{aligned} \widehat{\alpha}_{S.M} \xrightarrow{p} & \mathrm{E}(g(X)) - \beta\sqrt{\frac{\mathrm{var}(X^2)}{\mathrm{var}(X^2)+G}}\left\{\mathrm{E}(X^2) + \frac{1+\zeta}{m}\mathrm{E}(g(X))\right\} \\ = & \alpha + \beta\mathrm{E}(X^2)\left\{1 - \sqrt{\frac{\mathrm{var}(X^2)}{\mathrm{var}(X^2)+G}}\right\} - \frac{(1+\zeta)\beta}{m}\mathrm{E}(g(X))\sqrt{\frac{\mathrm{var}(X^2)}{\mathrm{var}(X^2)+G}} \\ = & \alpha + F. \end{aligned}$$

Using facts 4, 8 and 10,

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n} W_{b,i}^2(\zeta)S_i - (n^{-1}\sum_{i=1}^{n} W_{b,i}^2(\zeta))(n^{-1}\sum_{i=1}^{n} S_i) \\ \xrightarrow{p} & \mathrm{E}(X^2 g(X)) + \left\{\frac{1+\zeta}{m} + \frac{2\zeta}{m(m-1)}\right\}\mathrm{E}(g^2(X)) \\ & - \left\{\mathrm{E}(X^2) + \frac{1+\zeta}{m}\mathrm{E}(g(X))\right\}\mathrm{E}(g(X)) \\ = & \beta\mathrm{var}(X^2) + \frac{1+\zeta}{m}\beta^2\mathrm{var}(X^2) + \frac{2\zeta}{m(m-1)}\mathrm{E}(g^2(X)). \end{aligned} \tag{15}$$

Combining (15) and (14) leads to the result for $\widehat{\beta}_{S.M}$. Result for $\widehat{\alpha}_{S.M}$ is straightforward.

Proof of Lemma 6

We only need to show the results with $j = 1$. Using facts 11 and 14,

$$\frac{1}{3n}\sum_{i=1}^{n}(S_i^{(1)})^2 - (n^{-1}\sum_{i=1}^{n} S_i^{(1)})^2 \xrightarrow{p} (1 + \frac{1+\zeta}{m-1})^2\mathrm{E}(g^2(X)) - (\mathrm{E}(g(X)))^2 = \beta^2\mathrm{var}(X^2) + (1+\zeta)J. \tag{16}$$

Using facts 8 and 9,

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n}(W_{b,i}^{(j)}(\zeta))^4 - (n^{-1}\sum_{i=1}^{n}(W_{b,i}^{(j)}(\zeta))^2)^2 \\ \xrightarrow{p} & \mathrm{E}(X^4) + \frac{6(1+\zeta)}{m-1}\mathrm{E}(X^2 g(X)) + \frac{3(1+\zeta)^2}{(m-1)^2}\mathrm{E}(g^2(X)) - \left\{\mathrm{E}(X^2) + \frac{1+\zeta}{m-1}\mathrm{E}(g(X))\right\}^2 \\ = & \mathrm{var}(X^2) + (1+\zeta)K. \end{aligned} \tag{17}$$

Combining (16) and (17) leads to the result of $\widehat{\beta}_{PS.M}$. Now

$$\begin{aligned}\widehat{\alpha}_{PS.M} &\xrightarrow{p} \mathrm{E}(g(X)) + \frac{1+\zeta}{m-1}\mathrm{E}(g(X)) - \beta\sqrt{\frac{\mathrm{var}(X^2)+(1+\zeta)J}{\mathrm{var}(X^2)+(1+\zeta)K}}\left\{\mathrm{E}(X^2) + \frac{1+\zeta}{m-1}\mathrm{E}(g(X))\right\} \\ &= \alpha + I.\end{aligned}$$

Using facts 8, 11 and 12,

$$\begin{aligned}& n^{-1}\sum_{i=1}^{n}(W_{b,i}^{(1)}(\zeta))^2 S_i^{(1)} - (n^{-1}\sum_{i=1}^{n}(W_{b,i}^{(1)}(\zeta))^2)(n^{-1}\sum_{i=1}^{n} S_i^{(1)}) \\ \xrightarrow{p}\ & \mathrm{E}(X^2 g(X)) + \frac{1+\zeta}{m-1}\left\{\mathrm{E}(X^2 g(X)) + \mathrm{E}(g^2(X))\right\} + \frac{3(1+\zeta)^2}{(m-1)^2}\mathrm{E}(g^2(X)) \\ & - \left\{\mathrm{E}(X^2) + \frac{1+\zeta}{m-1}\mathrm{E}(g(X))\right\}\left\{1 + \frac{1+\zeta}{m-1}\right\}\mathrm{E}(g(X)) \\ =\ & \beta\mathrm{var}(X^2) + \frac{1+\zeta}{m-1}\left\{\mathrm{E}(X^2 g(X)) - \mathrm{E}(X^2)\mathrm{E}(g(X)) + \mathrm{E}(g^2(X)) - (\mathrm{E}(g(X)))^2\right\} \\ & + \frac{(1+\zeta)}{(m-1)^2}\left\{3\mathrm{E}(g^2(X)) - (\mathrm{E}(g(X)))^2\right\} \\ =\ & \beta\mathrm{var}(X^2) + (1+\zeta)M. \end{aligned} \tag{18}$$

Combining (18) and (17) leads to the result of $\widehat{\beta}_{PS.R}$. Now

$$\begin{aligned}& \widehat{\alpha}_{PS.R} \\ \xrightarrow{p}\ & \mathrm{E}(g(X)) + \frac{1+\zeta}{m-1}\mathrm{E}(g(X)) - \frac{\beta\mathrm{var}(X^2)+(1+\zeta)M}{\mathrm{var}(X^2)+(1+\zeta)K}\left\{\mathrm{E}(X^2) + \frac{1+\zeta}{m-1}\mathrm{E}(g(X))\right\} \\ =\ & \alpha + \mathrm{E}(X^2)\left\{\beta - \frac{\beta\mathrm{var}(X^2)+(1+\zeta)M}{\mathrm{var}(X^2)+(1+\zeta)K}\right\} + \frac{1+\zeta}{m-1}\mathrm{E}(g(X))\left\{1 - \frac{\beta\mathrm{var}(X^2)+(1+\zeta)M}{\mathrm{var}(X^2)+(1+\zeta)K}\right\} \\ =\ & \alpha + L.\end{aligned}$$

Proof of Theorem 3

Assume the true value of $\boldsymbol{\theta}$ is $\boldsymbol{\theta}_0$. Using (12) we have

$$0 = \sum_{i=1}^{n} S_{eff}^{*}(\mathbf{Y}_i, \widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} S_{eff}^{*}(\mathbf{Y}_i, \boldsymbol{\theta}_0) + \sum_{i=1}^{n}\frac{\partial S_{eff}^{*}(\mathbf{Y}_i, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^T}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}^*$ is in between $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}$. Therefore,

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{n^{-1}\sum_{i=1}^{n}\frac{\partial S_{eff}^{*}(\mathbf{Y}_i, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^T}\right\}^{-1} S_{eff}^{*}(\mathbf{Y}_i, \boldsymbol{\theta}_0) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[E\left\{\frac{\partial S_{eff}^{*}(\mathbf{Y}_i, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^T}\right\} + o_p(1)\right]^{-1} S_{eff}^{*}(\mathbf{Y}_i, \boldsymbol{\theta}_0).\end{aligned}$$

Because of (11), $E\left\{S^*_{eff}(\mathbf{Y}_i, \boldsymbol{\theta}_0)\right\} = 0$, hence

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[E\left\{ \frac{\partial S^*_{eff}(\mathbf{Y}_i, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^T} \right\} \right]^{-1} S^*_{eff}(\mathbf{Y}_i, \boldsymbol{\theta}_0) + o_p(1),$$

consequently, $\widehat{\theta}$ is a root-n consistent, asymptotically normal estimator of θ_0, with variance given as $A^{-1}BA^{-T}$ in Theorem 3. When $\eta^* = \eta_0$,

$$\begin{aligned} A &= -E\left\{ \frac{\partial S_{eff}(\mathbf{Y}, \theta_0)}{\partial \theta^T} \right\} = -\int \frac{\partial S_{eff}(\mathbf{Y}, \theta_0)}{\partial \theta^T} p_{\mathbf{Y}}(\mathbf{Y}, \theta_0) d\mu(\mathbf{Y}) \\ &= \int S_{eff}(\mathbf{Y}, \theta_0) \frac{\partial p_{\mathbf{Y}}(\mathbf{Y}, \theta)}{\partial \boldsymbol{\theta}^T}|_{\boldsymbol{\theta}_0} d\mu(\mathbf{Y}) = \int S_{eff}(\mathbf{Y}, \theta_0) \frac{\partial \log p_{\mathbf{Y}}(\mathbf{Y}, \theta)}{\partial \boldsymbol{\theta}^T}|_{\boldsymbol{\theta}_0} p_{\mathbf{Y}}(\mathbf{Y}, \theta_0) d\mu(\mathbf{Y}) \\ &= E(S_{eff} S^T_{\boldsymbol{\theta}}) = E(S_{eff} S^T_{eff}) = B, \end{aligned}$$

because S_{eff} is the orthogonal projection of $S_{\boldsymbol{\theta}}$.