Exact and Approximate Variance-Stabilizing Transformations for Two-Color Microarrays

B. Durbin, Department of Statistics, UC Davis, Davis, CA 95616*
D.M. Rocke, Department of Applied Science, UC Davis, Davis, CA, 95616

October 2, 2002

Abstract

Motivation. Durbin et al (2002), Huber et al (2002) and Munson (2001) independently introduced a family of transformations (the generalized-log family) which stabilizes the variance of microarray data up to the first order. However, for data from two-color arrays, tests for differential expression require that the variance of the difference of transformed observations be constant, rather than that of the transformed observations themselves.

Results. We introduce a transformation within the generalized-log family which stabilizes, to the first order, the variance of the difference of transformed observations. We also introduce transformations from the "started-log" and log-linear-hybrid families which provide good approximate variance stabilization of differences. Examples using control-control data show that any of these transformations may provide sufficient variance stabilization for practical applications, and all perform well compared to log ratios.

Contact. `bpdurbin@wald.ucdavis.edu`

Keywords. cDNA array, microarray, statistical analysis, transformation, normalization.

1 Introduction

Many traditional statistical methodologies, such as regression or ANOVA, are based on the assumptions that the data are normally distributed (or at least symmetrically distributed) with constant variance not depending on the mean of the data. If these assumptions are violated, the statistician may choose either to develop some new statistical technique which accounts for the specific ways

*To whom correspondence should be addressed.

in which the data fail to comply with the assumptions, or to transform the data. Where possible, data transformation is generally the easier of these two options (see Box and Cox, 1964, and Atkinson, 1985).

Data from gene-expression microarrays, which allow measurement of the expression of thousands of genes simultaneously, can yield invaluable information about biology through statistical analysis. However, microarray data fail rather dramatically to conform to the canonical assumptions required for analysis by standard techniques. Rocke and Durbin (2001) demonstrate that the measured expression levels from microarray data can be modeled as

$$y = \alpha + \mu e^{\eta} + \varepsilon \tag{1}$$

where y is the measured raw expression level for a single color, α is the mean background noise, μ is the true expression level, and η and ε are normally-distributed error terms with mean 0 and variance σ_η^2 and σ_ε^2 , respectively. The variance of y under this model is

$$\mathrm{Var}(y) = \mu^2 S_\eta^2 + \sigma_\varepsilon^2, \tag{2}$$

where $S_\eta^2 = e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1)$. Note that the variance is a quadratic function of the true expression μ.

The error structure of two-color spotted cDNA arrays can be modeled by an extended version of (1). With this microarray technology, mRNA from two different biological samples is reverse transcribed and labelled with two different fluorescent dyes, usually Cy3 and Cy5. The two samples are then hybridized to the same spotted cDNA array, resulting in two correlated measurements for each spot.

Rocke and Durbin (2001) model this pair of treatment and control observations for a single spot as

$$\begin{aligned} y_T &= \alpha_T + \mu_T e^{\eta_S + \eta_T} + \varepsilon_S + \varepsilon_T \\ y_C &= \alpha_C + \mu_C e^{\eta_S + \eta_C} + \varepsilon_S + \varepsilon_C \end{aligned}$$

where y_T and y_C are the raw signal intensities for the control and treatment samples, respectively, μ_T and μ_C are the true expression levels of the gene in question, η_S and ε_S are spot-specific multiplicative and additive error terms shared by y_T and y_C , and η_T, η_C, ε_T, and ε_C are multiplicative and additive error terms unique to control and treatment. Each error term is assumed to have mean 0 and to be stochastically independent from the others, with its own variance.

For the purposes of the following discussion, it will be more convenient to work with z_T and z_C rather than with y_T and y_C , where $z_T = y_T - \hat{\alpha}_T$. This presumes that any requisite background correction and normalization have already been applied to the data so that, to the first order, $\mathrm{E}(z_T) = \mu_T$ and $\mathrm{E}(z_C) = \mu_C$. The specific method normalization method used is left to the discretion of the reader.

A central question posed by a two-color microarray experiment is that of which genes are differentially expressed between the control and treatment samples. A common approach to determining differential expression is to examine the ratio z_T/z_C, or its logarithm $\ln(z_T/z_C)$. However, calculations based on the two-component model (1) and examination of mean-variance plots of microarray data show that $\ln(z)$ has a greatly inflated variance for μ close to 0 (Durbin et al., 2002). Due to this nonconstancy of variance, a log ratio that is statistically significant for one pair of true expression values (μ_T, μ_C) may not be significant for a different pair of values, even if the log ratio itself remains the same. Therefore, log ratios do not appear to provide an optimal means of determining differential expression.

The identity $\ln(z_T/z_C) = \ln(z_T) - \ln(z_C)$ does, however, suggest that one might approach the issue of differential expression using differences of transformations applied to treatment and control. We examine the behavior of

$$\Delta h(z_T, z_C) = h(z_T) - h(z_C)$$

for three different families of transformations: the generalized-log transformation of Durbin et al. (2002), Huber et al. (2002), and Munson (2001), the "started log" (Tukey, 1964, 1977), and the log-linear hybrid (Holder et al. 2001). Extending our previous work on one-color arrays (Rocke and Durbin 2002), we will discuss which member of each of these families provides the best variance stabilization for Δh.

2 The Generalized-Log Transformation

Durbin et al. (2002), Huber et al. (2002), and Munson (2001) independently discovered the application to gene-expression microarray data of transformation which stabilizes, to the first order, the variance of a random variable z satisfying

$$\mathrm{Var}(z) = a^2 + b^2\mu^2,$$

where $\mu = \mathrm{E}(z)$. This transformation may be written in several equivalent forms, but we will use

$$h_{\lambda_0} = \ln\left(\frac{z + \sqrt{z^2 + \lambda_0}}{2}\right) \tag{3}$$

where $\lambda_0 = a^2/b^2$. This transformation converges to $\ln(z)$ for large z, and is approximately linear at 0 (Durbin et al., 2002). The transformation and its inverse are monotonic functions with derivatives of all orders. Because its behavior for large values of μ is identical with the natural logarithm, and following Munson (2001), we will call this transformation a *generalized logarithm*.

Since there exist transformations of the family $h_\lambda(z) = \ln((z + \sqrt{z^2 + \lambda})/2)$ which stabilize the variances of z_T and z_C individually, it seems reasonable to search within this family for a transformation $h_\lambda(\cdot)$ such that $\Delta h_\lambda(z_T, z_C)$ has constant variance. For the purposes of testing for differential expression, we need to know the variance of a test statistic in the null case, that is, no

differential expression. Therefore, for each of these families of transformations we will focus on the behavior of $\mathrm{Var}(\Delta h(z_T, z_C))$ when $\mu_T = \mu_C = \mu$.

The asymptotic variance of $\Delta h_\lambda(z_T, z_C)$ for an unspecified parameter λ may be determined using the multivariate delta method. Once we have calculated the asymptotic variance of a function of λ, we may solve for lambda such that $\mathrm{AV}_{\mu_C=\mu_T=\mu}(\Delta h_\lambda(z_T, z_C))$ does not vary with μ. (We will adopt the notation $\mathrm{AV}(X)$ to denote the delta-method asymptotic variance of a random variable X.)

The details of the multivariate delta method are as follows: If $\mathbf{X}$ is a p-dimensional vector-valued random variable taking values near a constant vector θ, and $h(\cdot)$ is a smooth function mapping p-dimensional vectors into the real line, then the variance of $h(\mathbf{X})$ is, to the first order

$$\mathrm{AV}(\mathbf{X}) = \dot{h}(\theta)^\top \Sigma \, \dot{h}(\theta),$$

where Σ is the variance-covariance matrix of $\mathbf{X}$ and

$$\dot{h}(\mathbf{t}) = [\frac{\partial h}{\partial t_1}, \ldots, \frac{\partial h}{\partial t_p}]^\top.$$

Notice that $\Delta h_\lambda(z_T, z_C)$ is a function of the six independent random variables η_S, η_T, η_C, ε_S, ε_T, and ε_C, which have mean 0 and joint variance-covariance matrix $\Sigma = \mathrm{diag}(\sigma^2_{\eta_S}, \sigma^2_{\eta_T}, \sigma^2_{\eta_C}, \sigma^2_{\varepsilon_S}, \sigma^2_{\varepsilon_T}, \sigma^2_{\varepsilon_C})$, so for the following calculations we will use $\mathbf{X} = [\eta_S, \eta_T, \eta_C, \varepsilon_S, \varepsilon_T, \varepsilon_C]^\top$ and $\theta = \mathbf{0}$.

Using this technique we find that

$$\mathrm{AV}(\Delta h_\lambda(z_T, z_C)) = \frac{\mu^2(\sigma^2_{\eta_T} + \sigma^2_{\eta_C}) + \sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C}}{\mu^2 + \lambda}. \tag{4}$$

At $\mu = 0$ this becomes $(\sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C})/\lambda$, and as $\mu \to \infty$, $\mathrm{AV}(\Delta h_\lambda(z_T, z_C)) \to \sigma^2_{\eta_T} + \sigma^2_{\eta_C}$. If the variance is to be constant, at the very least it should be equal at $\mu = 0$ and as $\mu \to \infty$. Setting

$$\frac{\sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C}}{\lambda} = \sigma^2_{\eta_T} + \sigma^2_{\eta_C}$$

and solving for λ yields the candidate value

$$\lambda^\star = \frac{\sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C}}{\sigma^2_{\eta_T} + \sigma^2_{\eta_C}} \tag{5}$$

Inserting this value into (4) we find that

$$\mathrm{AV}(\Delta h_{\lambda^\star}(z_T, z_C)) = \sigma^2_{\eta_T} + \sigma^2_{\eta_C}, \tag{6}$$

which does not depend on μ. This member of the family of transformations $h_\lambda(z) = \ln((z+\sqrt{z^2+\lambda})/2)$ exactly stabilizes the asymptotic variance of $h_\lambda(z_T) - h_\lambda(z_C)$, allowing meaningful hypothesis tests to be performed on the differences. One may compare (5) with the expression for one color arrays of the optimal transformation parameter $\lambda = \sigma^2_\varepsilon/\sigma^2_\eta$ (Durbin et al. 2002).

3 The Started-Log Transformation

While the generalized-log transformation of section 2 of course provides the most exact variance stabilization, it may occasionally prove more convenient to use a transformation which only approximately stabilizes the variance of the difference of transformed observations. In particular, log ratios are occasionally touted as providing better interpretability than alternatives, despite their inherent problems with inflation of the variance of low-level observations. However, one problem with log ratios which is more difficult to ignore is that of negative observations. When μ_T or μ_T is near 0, z_T or z_C will often be negative, in which case the log ratio is not defined. An ad hoc solution is to simply discard data for which z_T or z_C is less than zero; however, this approach can result in the loss of valuable biological information.

Should one insist on using log ratios to determine differential expression, a modified version of the logarithm, called the "started logarithm" by Tukey (1964, 1977), can mitigate some of the problems with negative observations. This transformation takes the form

$$h_c(z) = \ln(z + c), \tag{7}$$

where $c > 0$. The delta-method variance of

$$\begin{aligned} \Delta h_c(z_T, z_C) &= h_c(z_T) - h_c(z_C) \\ &= \ln(\frac{z_T + c}{z_C + c}) \end{aligned}$$

under the null hypothesis $\mu_C + \mu_T = \mu$ is

$$\begin{aligned} \mathrm{AV}_{\mu_T=\mu_C=\mu}(\Delta h_c(z_T, z_C)) &= \frac{\mu^2(\sigma^2_{\eta_C} + \sigma^2_{\eta_T}) + \sigma^2_{\varepsilon_C} + \sigma^2_{\varepsilon_T}}{(\mu + c)^2} &(8)\\ &= \frac{q^2 + \mu^2 r^2}{(\mu + c)^2} &(9) \end{aligned}$$

where $q = \sqrt{\sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C}}$ and $r = \sqrt{\sigma^2_{\eta_T} + \sigma^2_{\eta_C}}$.

While no member of this family will exactly stabilize the asymptotic variance, we may ask for the choice of c which minimizes the maximum deviation of the variance from constancy. As $\mu \to \infty$, $\mathrm{AV}(\Delta h_c(z_T, z_C)) \to r^2$, which does not depend on c, so we will focus on the deviation of the variance from this limiting value.

At $\mu = 0$, $\mathrm{AV}(\Delta h_c(z_T, z_C)) = q^2/c^2$. The derivative of (9) with respect to μ is

$$\frac{2r^2\mu(\mu + c) + 2(q^2 + r^2\mu^2)}{(\mu + c)^3} \tag{10}$$

and the value of the derivative at $\mu = 0$ is $-2q^2/r^3 < 0$, indicating that the asymptotic variance decreases initially as μ increases away from 0. The denominator of (10) will never be negative for $\mu \geq 0$, so any change in the sign of the

derivative will occur where

$$2r^2\mu(\mu + c) - 2(q^2 + r^2\mu^2) = 0, \tag{11}$$

i.e., when $\mu = q^2/(r^2c)$. So, at $\mu = q^2/(r^2c)$, the asymptotic variance stops decreasing and starts increasing toward r^2.

The value of c that minimizes the maximum deviation from constancy will occur when the asymptotic variance at 0 is as much above the limiting value r^2 as the asymptotic variance at the minimum is below r^2. The asymptotic variance at the minimum is

$$\frac{q^2 + (q^2/(r^2c))^2 r^2}{(q^2/(r^2c) + c)^2} = \frac{q^2 r^2}{q^2 + r^2 c^2}.$$

In order to minimize the maximum deviation of the variance from constancy, we set the deviation at 0 equal to that at the minimum and solve for c, which yields

$$\frac{q^2}{c^2} - r^2 = r^2 - \frac{q^2 r^2}{q^2 + r^2 c^2},$$

or

$$c = q/(2^{\frac{1}{4}} r)$$

. The minimized maximum deviation of the variance from constancy is

$$\frac{q^2}{c^2} - r^2 = r^2\sqrt{2} - r^2$$

and the ratio of the standard deviation at 0, which is $2^{\frac{1}{4}} r$, to the limiting standard deviation r is about 1.2

We illustrate the behavior of this transformation using an example from Rocke and Durbin (2001). The parameter values estimated from the data of Bartosiewicz et al. (2000) are shown in Table 1. For these parameter values, $q = 7460$, $r = 0.0674$, and the optimal shift constant c for the started log is 93,200. Figure 1 shows the standard deviation function of $\Delta h_c(z_T, z_C) = \ln(z_T + c) - \ln(z_C + c)$ for the optimal shift constant, as well as for two other values. The horizontal line on the plot shows the limiting standard deviation $r = 0.0674$.

The uppermost line in Figure 1 shows the standard deviation function when $c = 0$, corresponding to the log ratio $\ln(\frac{z_T}{z_C})$. The standard deviation of the log ratio approaches infinity as μ approaches 0, but decreases towards r as μ increases. The middle line shows the standard deviation function for $c = 25000$, which would correspond roughly to taking the log ratio of data that had been adjusted so that $\alpha_T = \alpha_C$, but without subtracting the expression background. The standard deviation at $\mu = 0$ for the transformation using $c = 25000$ is 0.299 and the minimum standard deviation is 0.0657, yielding a maximum deviation from r of 0.231. The lowest line on the plot shows the standard deviation function for the optimal transformation using $c = 93200$. For this constant, the standard deviation at $\mu = 0$ is 0.0801 and the minimum standard deviation

is 0.0516, yielding a maximum deviation from r of 0.0127. These results are summarized in Table 2, which shows the standard deviation at $\mu = 0$, the argmin (value where the minimum occurs), the minimum standard deviation, and the maximum deviation of the standard deviation from constancy for different values of c.

The optimal started log transformation seems to provide reasonable variance stabilization of the difference of transformed observations for the parameters given. According to this theoretical plot, the log-ratios of the background-corrected data suffer from infinite variance as μ approaches 0. The log-ratios of the color-normalized, uncorrected data also do not perform as well as the optimal transformation, with a maximum deviation from constancy of the standard deviation more than 18 times of that for the optimal transformation.

4 The Log-Linear-Hybrid Transformation

A third class of transformations which may prove useful in the analysis of microarray data is the log-linear hybrid (Holder et al. 2001). According to the two-component model (1), for μ close to 0, the untransformed data have approximately constant variance, and for μ large, $\ln(z)$ has approximately constant variance (Rocke and Durbin, 2001). This suggests that we might use a linear transformation for small z and a log transformation for large z.

Let

$$h_k(z) = \begin{cases} c + dz, & z \le k \\ \ln(z), & z > k \end{cases} \tag{12}$$

If we choose c and d so that $h_k(z)$ is continuous with continuous derivative at k, we get $c = 1/k$ and $d = \ln(k) - 1$, yielding

$$h_k(z) = \begin{cases} z/k + \ln(k) - 1, & z \le k \\ \ln(z), & z > k \end{cases} \tag{13}$$

It remains to choose k to minimize the maximum deviation of the variance of

$$\Delta h_k(z_T, z_C) = h_k(z_T) - h_k(z_C) \tag{14}$$

from constancy. The delta method variance of (14) takes 4 different forms, depending on the values of z_T and z_C relative to the splice point k. Therefore,

under the null hypothesis $\mu_T = \mu_C = \mu$,

$$\mathrm{AV}(\Delta h_k(z_T, z_C)) = \begin{cases} \frac{\mu^2(\sigma^2_{\eta_T}+\sigma^2_{\eta_C})+\sigma^2_{\varepsilon_T}+\sigma^2_{\varepsilon_C}}{k^2}, & z_T, z_C \le k \\ \sigma^2_{\eta_T} + \sigma^2_{\eta_C} + \frac{\sigma^2_{\varepsilon_T}+\sigma^2_{\varepsilon_C}}{\mu^2}, & z_T, z_C > k \\ (1-\frac{\mu}{k})^2\sigma^2_{\eta_S} + \sigma^2_{\eta_T} + \frac{\mu^2}{k^2}\sigma^2_{\eta_C} & \\ \quad + (\frac{1}{\mu}-\frac{1}{k})^2\sigma^2_{\varepsilon_S} + \frac{\sigma^2_{\varepsilon_T}}{\mu^2} + \frac{\sigma^2_{\varepsilon_C}}{k^2}, & z_T > k, z_C \le k \\ (\frac{\mu}{k}-1)^2\sigma^2_{\eta_S} + \frac{\mu^2}{k^2}\sigma^2_{\eta_T} + \sigma^2_{\eta_C} & \\ \quad + (\frac{1}{k}-\frac{1}{\mu})^2\sigma^2_{\varepsilon_S} + \frac{\sigma^2_{\varepsilon_T}}{k^2} + \frac{\sigma^2_{\varepsilon_C}}{\mu^2}, & z_T \le k, z_C > k \end{cases} \tag{15}$$

When $\mu = 0$,

$$\begin{aligned} \mathrm{AV}(\Delta h_k(z_T, z_C)) &= \frac{\sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C}}{k^2} \\ &= \frac{q^2}{k^2}, \end{aligned}$$

where $q = \sqrt{\sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C}}$, as in section 3. As $\mu \to \infty$,

$$\begin{aligned} \mathrm{AV}(\Delta h_k(z_T, z_C)) &\to \sigma^2_{\eta_T} + \sigma^2_{\eta_C} \\ &= r^2, \end{aligned}$$

where $r = \sqrt{\sigma^2_{\eta_T} + \sigma^2_{\eta_C}}$, also as in section 3.

Notice that when $\mu = k$, all four expressions become

$$\sigma^2_{\eta_T} + \sigma^2_{\eta_C} + \frac{\sigma^2_{\varepsilon_T} + \sigma^2_{\varepsilon_C}}{k^2} = r^2 + \frac{q^2}{k^2}.$$

It can be seen that the value of k which minimizes the maximum deviation of the variance from constancy will be the one for which the variance at 0 is as much below the limiting value r^2 as the variance at the splice point is above r^2. Setting

$$r^2 - \frac{q^2}{k^2} = r^2 + \frac{q^2}{k^2} - r^2$$

yields

$$k = q\sqrt{2}/r$$

. With this value of k, the maximum deviation of the variance from constancy is $r^2/2$ and the ratio of the standard deviation of the difference at 0 to the limiting value r is about 0.7.

Figure 2 shows plots of the standard deviation function for the optimal log-linear-hybrid transformation, the optimal started-log transformation, and the optimal generalized logarithmic (variance-stabilizing) transformation with constant $\lambda = 1.23 \times 10^{10}$. The parameters used were those estimated from the data of Bartosiewicz et al. (2000) and are shown in Table 1. The optimal splice point

for the log-linear hybrid transformation is 157,000. For the purposes of plotting the standard deviation function for the log-linear-hybrid transformation, we assume z_T and z_C to both lie either above or below the splice point. For the log-linear hybrid transformation, $\sqrt{\mathrm{AV}\Delta h_k(z_T, z_C)} = 0.0476$ when $\mu = 0$ and $\sqrt{\mathrm{AV}\Delta h_k(z_T, z_C)} = 0.0825$ when $\mu = k$, resulting in a maximum deviation from the limiting value $r = 0.0674$ of 0.0197. The maximum deviation from constancy for the started log transformation is 0.0127, so the started log appears to behave better in this case than the log-linear hybrid, with the transformation of the family proposed by Durbin et al. (2002), Huber et al. (2002), and Munson (2001) of course providing the best variance stabilization. Since the maximum deviation from constancy of variance for the started log transformation is about $.41r^2$, and that for the log-linear hybrid is $.5r^2$, the log-linear hybrid always has the larger maximum deviation from constancy. However, as Figure 2 shows, differences of observations transformed using the log-linear hybrid transformation appear to reach constant variance much sooner than those transformed with the started log transformation. Any of these transformations may be sufficient to stabilize the variance of the difference of transformed observations for practical purposes.

5 Examples

We illustrate the performance of these transformations with additional data from Bartosiewicz et al. (2000). We will use a small subset of the data presented in that paper, featuring control vs. control experiments, in order to determine the behavior of the transformed data when there is no differential expression. For these data, two groups of three mice were each treated with .10 mg/kg of corn oil. mRNA from the livers of the mice was extracted, pooled, and reverse-transcribed into fluor-labelled cDNA, with one group labelled with Cy5 and one group labelled with Cy3. Notice that this is not true self-self data, since three different mice were used for each group. The cDNA was then hybridized to a spotted array in which each gene was replicated between 6 and 14 times.

Parameters for the two-component model were estimated as described in Rocke and Durbin (2001), yielding $\alpha_C = 0.353$, $\alpha_T = 0.139$, $\sigma_{\varepsilon C} = 0.335$, $\sigma_{\varepsilon T} = 0.0585$, $\sigma_{\varepsilon S} = 0.0747$, $\sigma_{\eta C} = 0$, $\sigma_{\eta T} = 0.135$, and $\sigma_{\eta S} = 0.143$. These model parameters yield the transformation parameters $\lambda = 6.33$ for the generalized-log transformation, $c = 2.12$ for the started-log transformation, and $k = 3.56$ for the log-linear-hybrid transformation.[1]

Figures 3–5 show the robustly-estimated replicate standard deviation of differences of transformed observations against the robustly-estimated mean for the generalized-log, started-log, and log-linear-hybrid transformations. The robust mean was estimated using the S-Plus function location.m, and the robust standard deviation was estimated using the S-Plus function scale.a. The solid

[1] Although these data come from the same paper as the data from which the parameters in section 3 were estimated, they were processed using a different scanner and different image-processing software, which accounts for the dramatic difference in scale.

line on each plus shows a lowess smooth fit to the robust means and standard deviations. The plots are centered to the left of 0, a phenomenon likely due to dye bias or to true differences between the two pooled groups of animals.

In each case, the standard deviation appears relatively constant when compared to the mean. Furthermore, the three plots look quite similar, indicating that each of these transformations does an adequate job of stabilizing the variance of the data. For comparison, Figure 6 shows the robustly-estimated replicate mean and standard deviation of the log ratios of the data. We removed the 86 negative numbers (out of a total sample size of 2,304) before taking the log transformation. The standard deviation spikes dramatically as the mean decreases. Any of the three optimal transformations presented above stabilizes the variance of the data much better than the log transformation.

6 Conclusions

We have presented three variance-stabilizing transformations for gene-expression microarray data from two-color arrays, one which exactly stabilizes the delta-method variance of differences of transformed observations, and two other transformations, the started-log and log-linear hybrid transformations, which provide approximate stabilization of the delta-method variance. When applied to actual data, each of these transformations appears to adequately stabilize the variance of differences of transformed observations, and all of these transformations provide much better variance stabilization than the log transformation.

7 References

Atkinson, A.C. (1985) *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis* Clarendon Press, Oxford.

Bartosiewicz, M., Trounstine, M., Barker, D., Johnston, R., and Buckpitt, A. (2000) Development of a toxicological gene array and quantitative assessment of this technology. *Archives of Biochemistry and Biophysics.*, 376, 66–73.

Box, G.E.P., and Cox, D.R. (1964) "An analysis of transformations," *Journal of the Royal Statistical Society, Series B (Methodological)*, **26**, 211–252.

Durbin, B.P., Hardin, J.S., Hawkins, D.M., and Rocke, D.M. (2002) "A variance-stabilizing transformation for gene-expression microarray data," *Bioinformatics*, **18**, S105–S110.

Hawkins, D.M. (2001) "Diagnostics for conformity of paired quantitative measurements," *Statistics in Medicine*, **21**, 1913–1935.

Holder, D., Raubertas, R.F., Pikounis, V.B., Svetnik, V., and Soper, K. (2001) "Statistical analysis of high density oligonucleotide arrars: A SAFER approach," GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data.

Huber, W., Von Heydebreck, A., Poustka, A., and Vingron, M. (2002) "Variance stabilization applied to microarray data calibration and to the quantification of differential expression," *Bioinformatics*, **18**, S96–S104.

Munson, P. (2001) "A 'Consistency' Test for Determining the Significance of Gene Expression Changes on Replicate Samples and Two Convenient Variance-stabilizing Transformations," GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data.

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

Rocke, D.M. and Durbin, B.P. (2002) "Approximate variance-stabilizing transformations for gene-expression microarray data," manuscript.

Tukey, J.W. (1964) "On the comparative anatomy of transformations," *Annals of Mathematical Statistics*, **28**, 602–632.

Tukey, J.W. (1977) *Exploratory Data Analysis* Reading, MA: Addison-Wesley.

α_C	25,300
α_T	24,800
σ_{ε_S}	5,270
σ_{ε_C}	0
σ_{ε_T}	7,460
σ_{η_S}	0.211
σ_{η_C}	0.0598
σ_{η_T}	0.0311
λ	1.23×10^{10}
q	7,460
r	.0674
optimal c	93,200
optimal k	157,000

Table 1: Parameters from the Two-Component Model for the Examples in Figures 1 and 2

c	S.D. at 0	Argmin	Min. S.D.	Max. Dev. S.D.
93,200	0.0801	132,000	0.0516	0.0127
0	∞	∞	0.0674	∞
25,000	0.299	491,000	0.0657	0.231

Table 2: Summary of the Standard Deviation Functions Shown in Figure 1

Figure 1: S.D. of the Difference of Started Logs for Three Values of the Constant

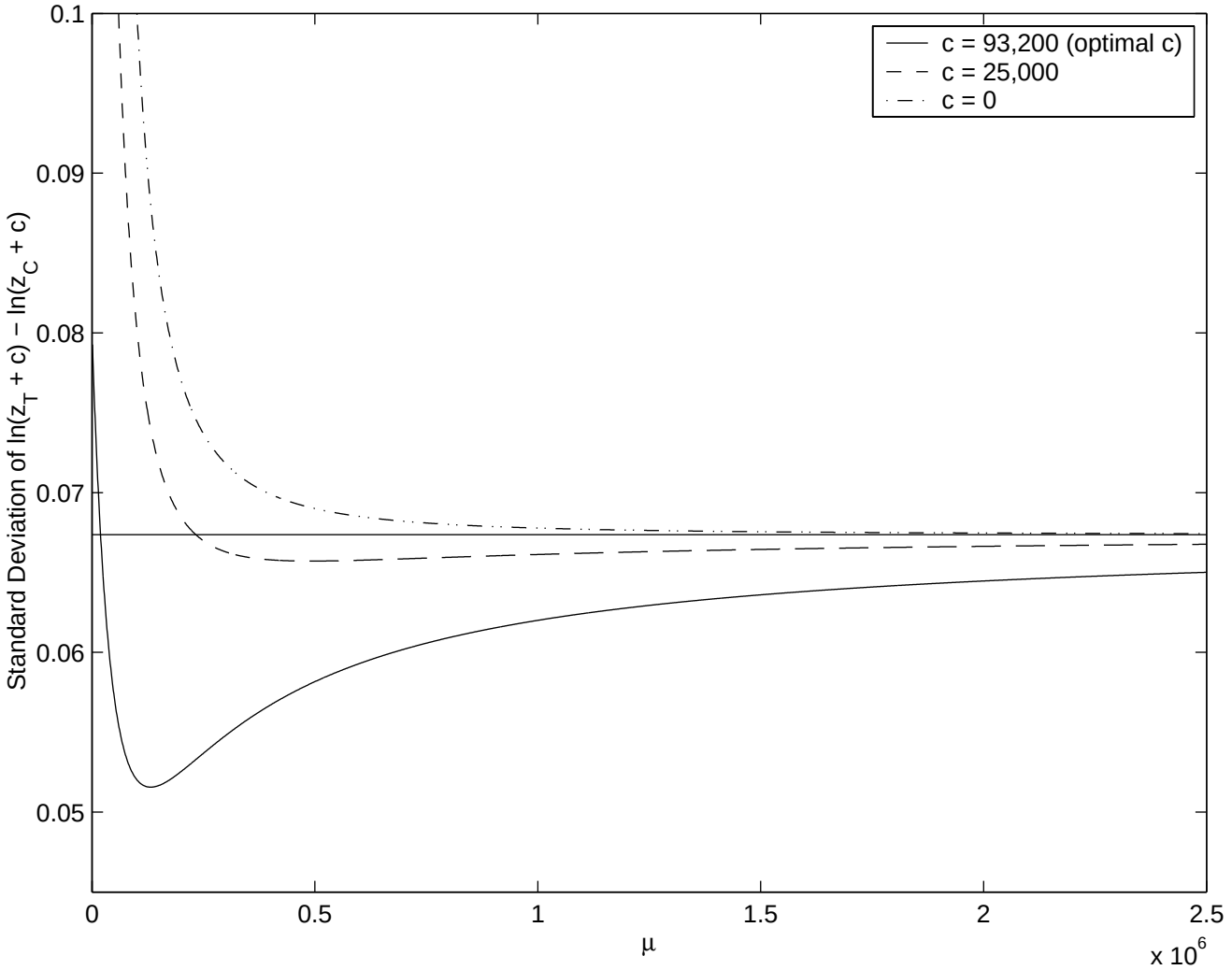

Figure 2: S.D. of Δ h for the Optimal Member of Three Families of Transformations

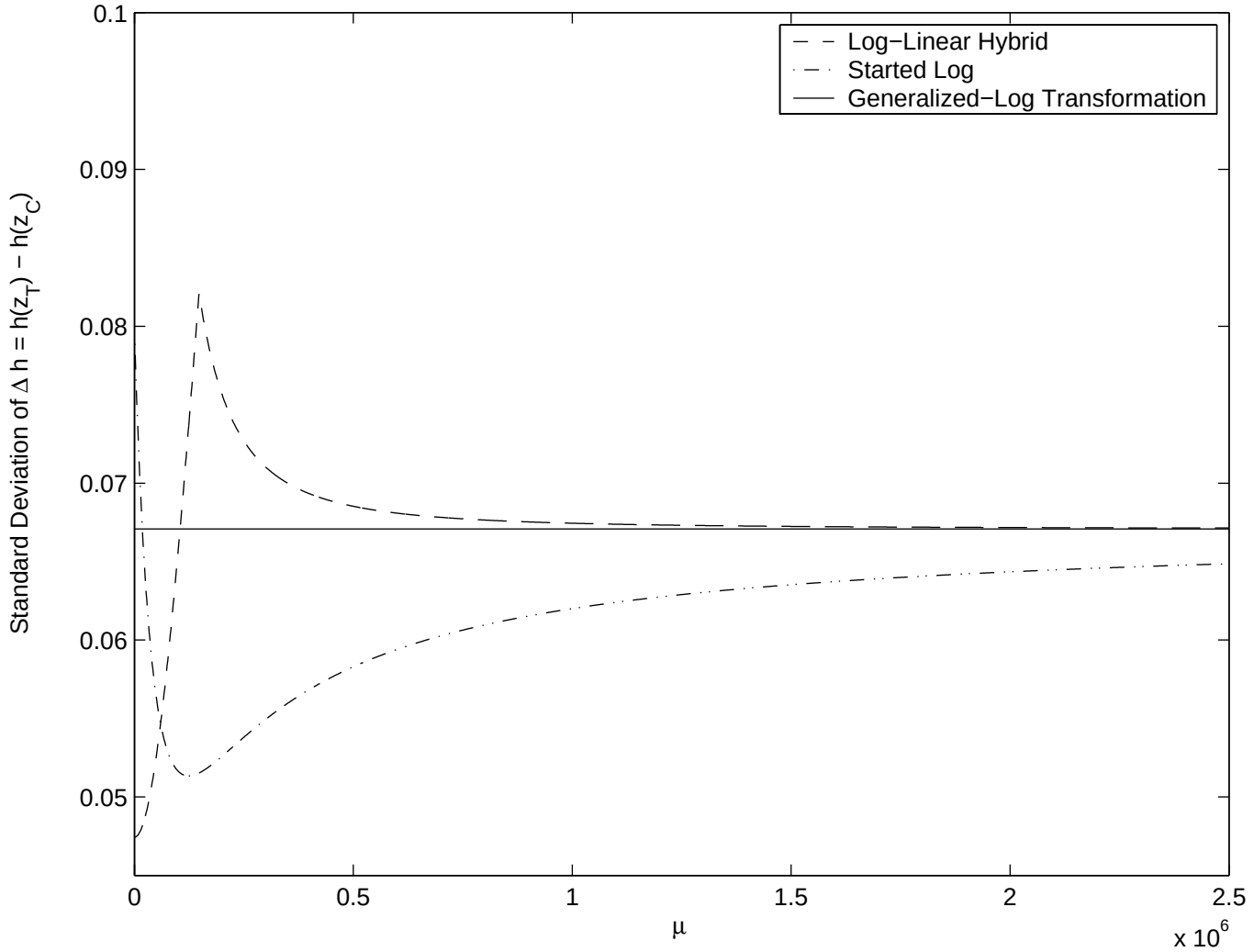

Figure 3: Robust Mean and Standard Deviation of Replicate Differences, Generalized−Log Transformation

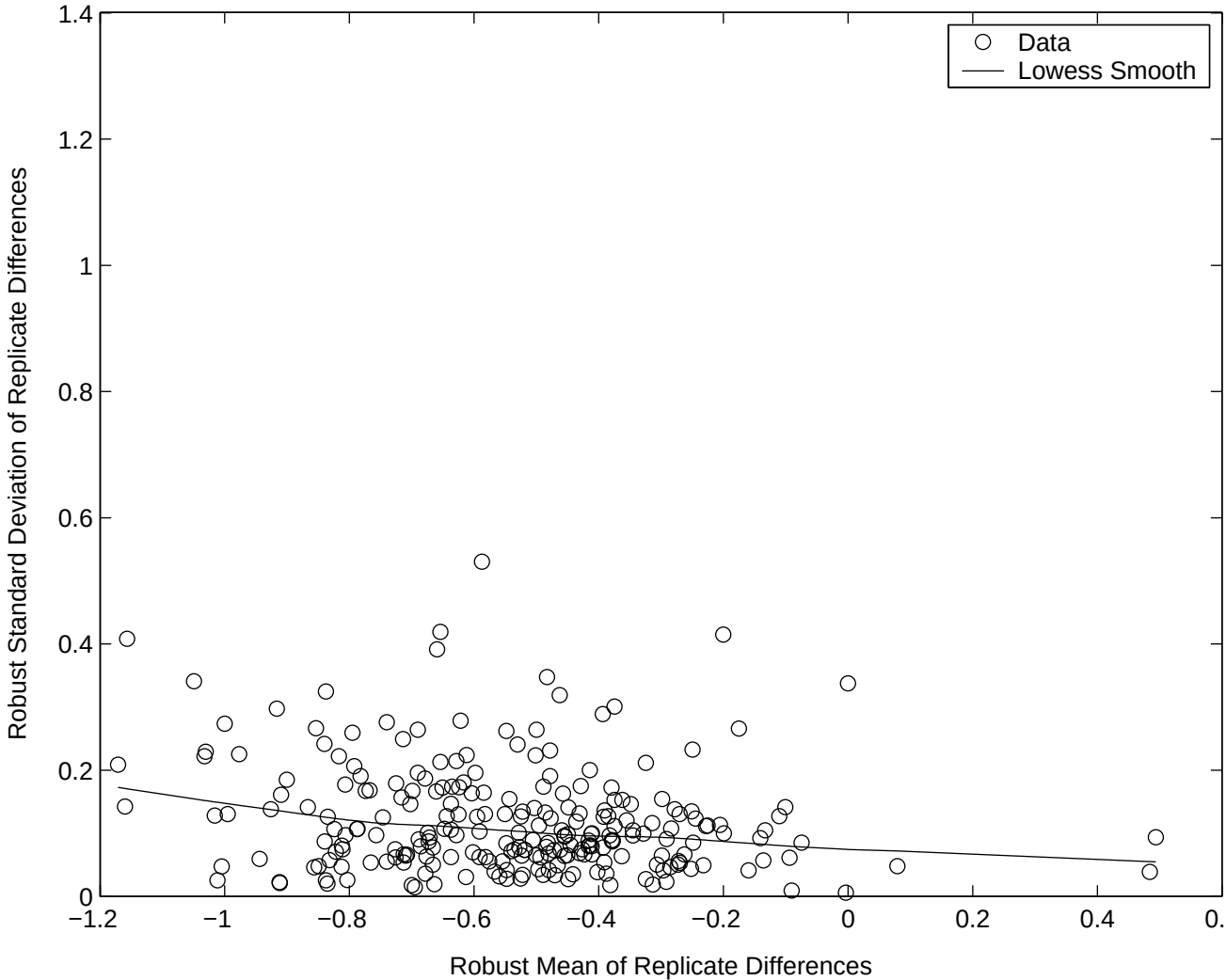

Figure 4: Robust Mean and Standard Deviation of Replicate Differences, Started−Log Transformation

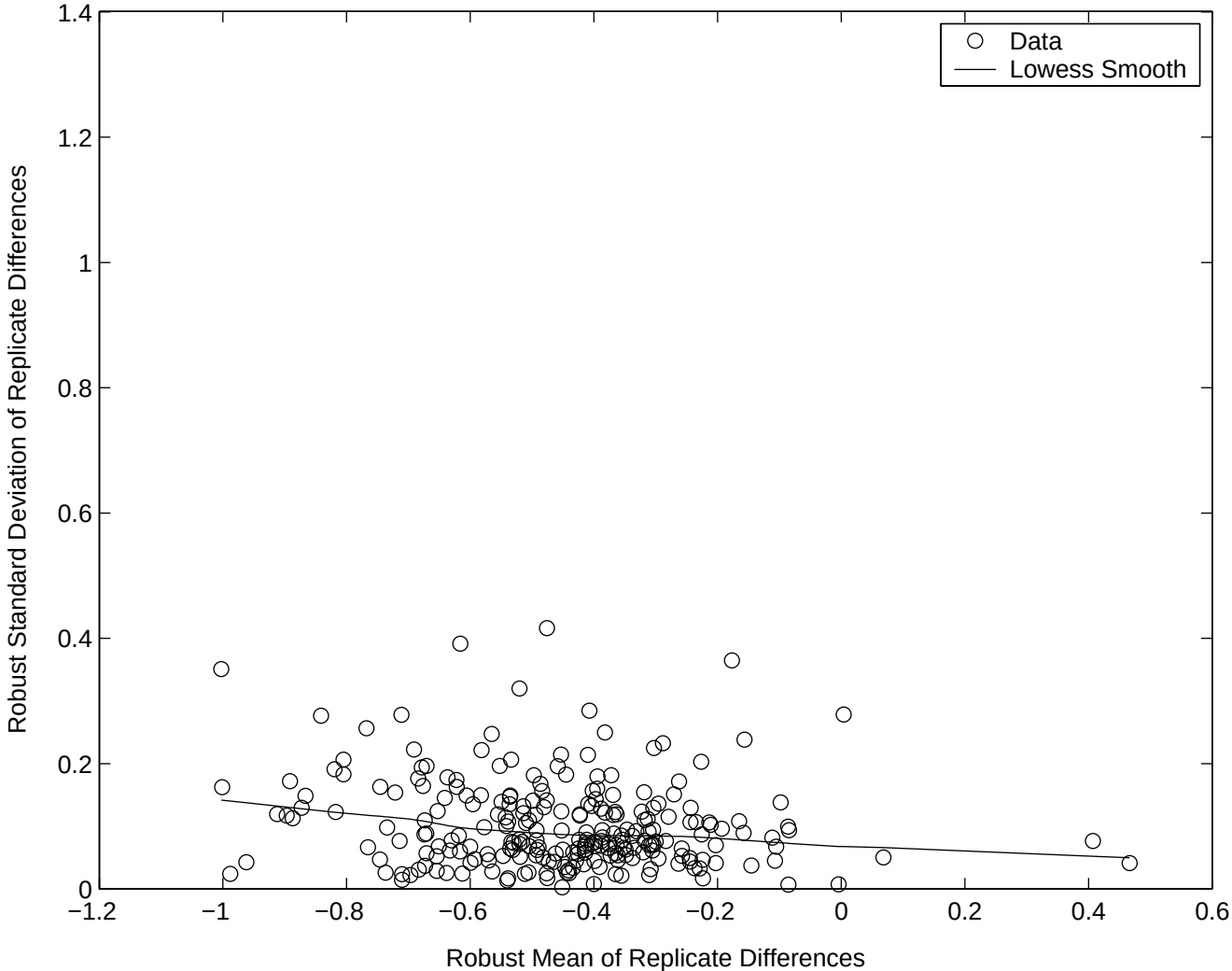

Figure 5: Robust Mean and Standard Deviation of Replicate Differences, Log−Linear−Hybrid Transformation

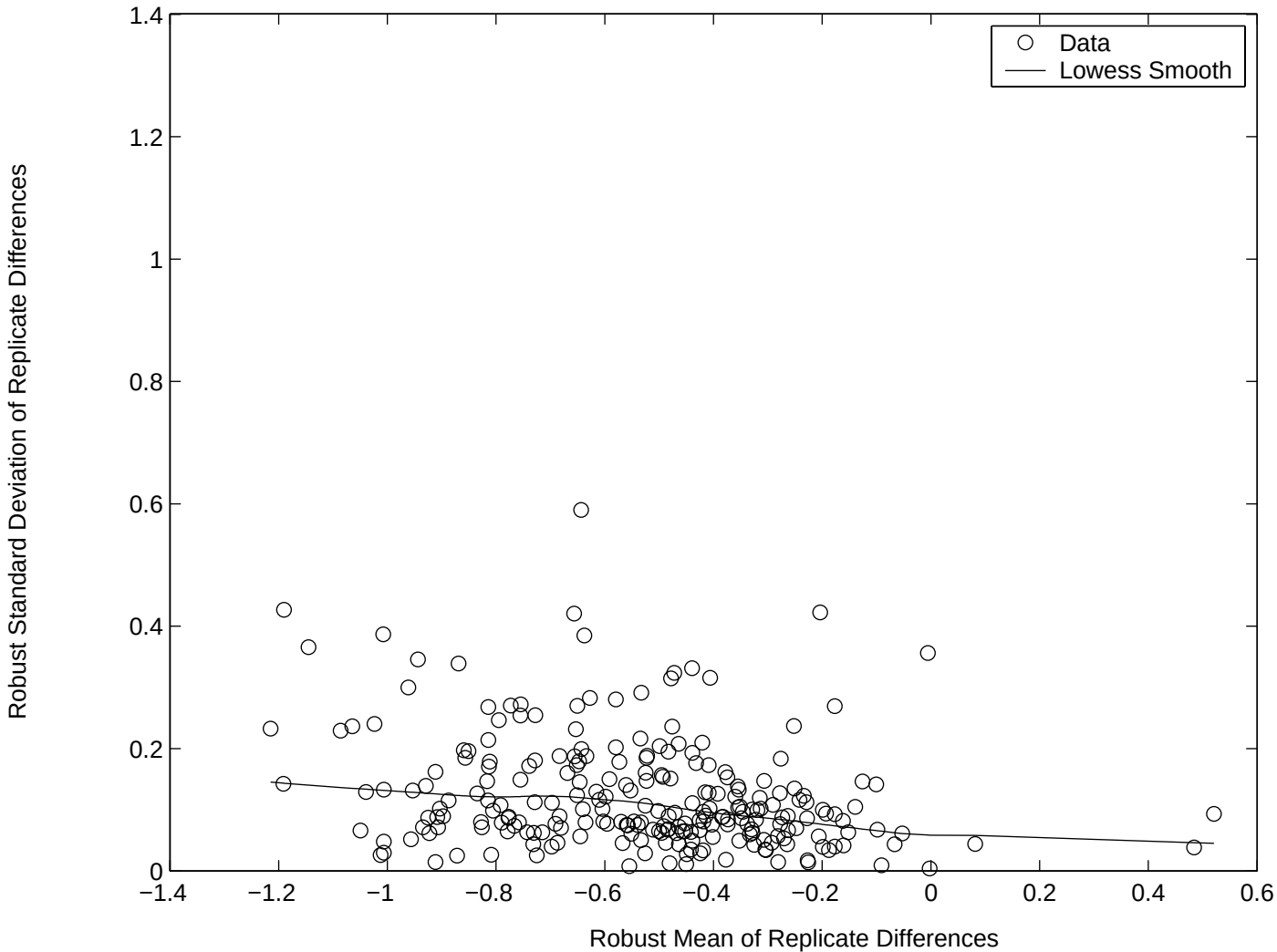

Figure 6: Robust Mean and Standard Deviation of Replicate Differences, Log Transformation