UC Davis
IDAV Publications

Title
Estimation of Transformation Parameters for Microarray Data

Permalink
https://escholarship.org/uc/item/0jq0t1p5

Journal
Bioinformatics, 19

Authors
Durbin, Blythe
Rocke, David

Publication Date
2003

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Vol. 19 no. 11 2003, pages 1360–1367
DOI: 10.1093/bioinformatics/btg178

Estimation of transformation parameters for microarray data

Blythe Durbin [1,*] and David M. Rocke [2]

[1]Department of Statistics and [2]Department of Applied Science, UC Davis, Davis, CA 95616, USA

Received on September 9, 2002; revised on February 19, 2003; accepted on February 21, 2003

ABSTRACT

Motivation and Results: 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. We introduce a method for estimating the transformation parameter in tandem with a linear model based on the procedure outlined in Box and Cox (1964). We also discuss means of finding transformations within the generalized-log family which are optimal under other criteria, such as minimum residual skewness and minimum mean-variance dependency.

Availability: R and Matlab code and test data are available from the authors on request.

Contact: bpdurbin@ucdavis.edu

1 INTRODUCTION

Gene-expression microarrays, with their ability to measure the expression of thousands of genes simultaneously, have the potential to revolutionize our understanding of the connection between an organism's genetic makeup and its phenotype. However, data from microarrays have proven surprisingly resistant to analysis by standard statistical techniques, which has somewhat slowed the rate at which new information has been gleaned from this technology. This is caused in part by the failure of microarray data to conform to the key assumptions on which many standard statistical techniques, such as linear regression and analysis of variance, are based. These techniques often require that one assume that data come from a normal distribution (or at least a symmetric distribution), that the data have a simple mean structure, and that the data have a constant error variance which does not depend on the mean of the data.

Violation of these assumptions can cause severe problems in statistical analysis of expression data. In the simplest setting, in which a single gene is compared in exactly two groups, much of the problem can be dealt with using the unequal-variance, two-sample t-test instead of the equal variance alternative, but in any more complex design than this, the assumption of equal variance is fundamental. For example, in the next simplest case of comparing more than two groups by one-way ANOVA, power is poor and size is not maintained if variance inhomogeneities occur. The problem is even more intractable for two-way and more complex designs.

When confronted with data that fail to conform to one or more of the standard assumptions, we may choose to address problems individually or to take a more global approach. For example, statistical weighting may be used to address the problem of nonconstant variance, but will fail to correct any skewness in the data. A transformation-based approach, on the other hand, can often correct multiple problems. We will compare the performance of a transformation intended to correct several problems simultaneously with those of transformations aiming to optimize a single criterion.

1.1 The generalized-log transformation

We will focus our attention on the family of generalized-log transformations, motivated by the two-component error model of Rocke and Durbin (2001). In this work it was demonstrated that the relationship between the true expression for an observation from a given gene and the measured expression can be modeled as

$$y = \alpha + \mu e^{\eta}(\eta) + \varepsilon \qquad (1)$$

where y is the measured expression for a single observation (either control or treatment in the case of a two-color array) for a given gene on a microarray, α is the mean expression background for the given array and sample, μ is the true expression, and η and ε are normally distributed error terms, with variances σ_η^2 and σ_ε^2, respectively. This model also works well for Affymetrix GeneChip arrays either applied to the PM-MM data or to individual probes.

Observations from the two-component model (1) have variance

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

*To whom correspondence should be addressed.

Bioinformatics 19(11) © Oxford University Press 2003; all rights reserved.

where $S_\eta^2 = e(\sigma_\eta^2)(e(\sigma_\eta^2) - 1)$. In Durbin *et al.* (2002), Huber *et al.* (2002) and Munson (2001) it was shown that for a random variable z satisfying $V(z) = a^2 + b^2\mu^2$, with $E(y) = \mu$, there is a transformation that stabilizes the variance to the first order. There are several equivalent ways of writing this transformation, but we will use

$$h_\lambda(z) = \ln(z + \sqrt{z^2 + \lambda}),$$

where $\lambda = a^2/b^2 = \sigma_\varepsilon^2/S_\eta^2$ and $z = y - \alpha$ or $y - \hat{\alpha}$. We shall refer to this transformation as the generalized-log transformation, as in Munson (2001), since the log transformation is a special case of this family for $\lambda = 0$.

(We will not specifically address background subtraction in this work; however, for proper application of the transformation the data must first have been adjusted so that $E(z) = \mu$, that is, the expectation of the adjusted data must be equal to the true expression level. In the cDNA array example of Section 2.1 a global expression background was subtracted from each channel prior to application of the transformation, 'global' meaning that all observations from the same channel and chip are assumed to share the same value of α following image processing. Background subtraction, in principle, can be accomplished as part of the transformation, with the caveat that the more parameters one includes in the transformation, the more difficult the estimation becomes, due to the increase in the dimension of the search space. In light of this, more sophisticated normalization methods, such as print-tip normalization, are best applied prior to data transformation.)

The generalized log transformation converges to $\ln(z) + \ln(2)$ for large z (equivalent to a log transformation, as the additive constant does not affect the strength of the transformation) , and is approximately linear at 0 (Durbin *et al.*, 2002). The inverse transformation is

$$h_\lambda^{-1}(u) = (e^u - \lambda e^{-u})/2.$$

Both h_λ and its inverse are monotonic functions, defined for all values of z and u, with derivatives of all orders.

When transforming data from two-color arrays or from complex multi-array experiments, the closed form expression for the transformation parameter shown in (1.1) is less useful than in the single color, single array case. Even data from different colors on the same two-color array might have different estimated values for the model parameters σ_η^2 and σ_ε^2, which makes it unclear exactly how we should obtain the transformation parameter. Pooling of data from different sources in order to estimate parameters could work well for some designs, but is not very flexible. An estimation method which specifically accounts for the structure of the data would be useful in these situations.

One such approach is to fit a linear model to the data while simultaneously estimating the transformation parameter via maximum likelihood, as was done in Box and Cox (1964). The linear model structure will allow us to account for the different sources of variation in the data, such as variation between arrays, between replicated spots on the same array, and between colors on the same array, in our estimation of the transformation parameter. Furthermore, the linear model, fit to appropriately transformed data, can itself be a useful analysis tool. An example of such an analysis would be the ANOVA normalization method for microarray data developed in Kerr *et al.* (2000).

2 MAXIMUM-LIKELIHOOD ESTIMATION

The maximum-likelihood estimation of the linear model and transformation parameters can be conducted essentially as in Box and Cox (1964), with the key distinction being that we shall search for an optimal transformation within the family of generalized log transformations, as in (1.1), rather than among the power transformations. We shall omit some of the details of the derivation; the interested reader may refer to Box and Cox (1964).

The procedure outlined in Box and Cox (1964) is as follows: Suppose that there exists some λ such that the transformed observations $\{h_{i,\lambda}\}$ have independent normal distributions with linear mean structure and constant variance σ^2. That is, suppose there exists lambda such that

$$\boldsymbol{h}_\lambda = (h_{1,\lambda}, \ldots, h_{n,\lambda})^\top = \boldsymbol{X}\beta + \varepsilon \tag{3}$$

where n is the number of observations in the data set, $\boldsymbol{X}$ is the design matrix from the linear model, β is a fixed vector of unknown linear model parameters, and $\varepsilon \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$.

The likelihood of the untransformed observations is therefore a normal likelihood in terms of the transformed observations times the Jacobian of the transformation (which allows for the change in scale of the data due to transformation). For the generalized-log transformation, the Jacobian is

$$J(z, \lambda) = \prod_{i=1}^{n} \left| \frac{dh_{i,\lambda}}{dz_i} \right| \tag{4}$$

$$= \prod_{i=1}^{n} 1/\sqrt{z_i^2 + \lambda}. \tag{5}$$

Box and Cox (1964) then suggest that if one should divide each transformed observation by the nth root of the Jacobian, the likelihood of the original data in terms of the Jacobian-corrected transformed data will be, approximately, a normal likelihood, rather than a normal likelihood times the Jacobian. (The 'approximate' nature of the likelihood comes from the fact that we are ignoring the variability of the nth root of the Jacobian, which will be quite minimal given the size of most microarray data sets).

Therefore, we will model the Jacobian-corrected transformed observations w_λ as

$$w_\lambda = X\beta + \varepsilon, \tag{6}$$

$$w_{i,\lambda} = h_{i,\lambda} \cdot \left(\prod_{i=1}^{n} \sqrt{z_i^2 + \lambda} \right)^{1/n}$$

and

$$h_{i,\lambda} = \ln\left(z_i + \sqrt{z_i^2 + \lambda} \right).$$

If we, for the moment, regard the transformation parameter as fixed (but still unknown), it is a relatively simple matter to obtain closed-form maximum-likelihood estimates of the linear model parameters β and σ^2 in terms of the unknown λ. These formulas may then be plugged into the log likelihood to obtain the partially-maximized log likelihood

$$l_{\max}(\lambda) = -\frac{n}{2} \ln \hat{\sigma}^2(\lambda)$$
$$= -\frac{n}{2} \mathrm{SSE}(z, \lambda)/n,$$

$$\mathrm{SSE}(z, \lambda) = \sum_{i=1}^{n} (w_i - \hat{w}_i)^2, \tag{7}$$

and $\hat{w}_i$ is the predicted value for the ith observation under the linear model fit to w_λ. The partially-maximized log likelihood depends on the data only through $\mathrm{SSE}(z, \lambda)$, and is a monotone decreasing function of this quantity, so we may find $\hat{\lambda}$, the MLE of λ, simply by minimizing $\mathrm{SSE}(z, \lambda)$ (Box and Cox, 1964). A minimum value may be found by plotting the error sum of squares as a function of λ, or via numerical optimization methods, such as Newton's method (see Appendix for details). Estimates of β and σ^2 on the scale of the transformed data without the Jacobian correction may be obtained by fitting the linear model again using the MLE, $\hat{\lambda}$, as the transformation parameter or by multiplying $\hat{\beta}$ by $J^{1/n}(z, \hat{\lambda})$ and multiplying $\hat{\sigma}^2$ by $J^{2/n}(z, \hat{\lambda})$.

Referring to a procedure for estimation of a transformation as a 'maximum-likelihood' method might cause some confusion with readers familiar with the maximum-likelihood method presented in Huber *et al.* (2002). One should bear in mind that 'maximum-likelihood' refers to a very large class of estimation methods, which are distinguished from one another primarily by the model for which parameters are estimated. The method presented above models the residuals from the linear model following transformation, whereas the method presented in Huber *et al.* (2002) models the unexpressed genes. Their method treats the differentially expressed genes as outliers, which are disregarded by the robust estimation procedure, and relies heavily on the assumption that differentially expressed genes constitute only a small fraction of the data.

2.1 Examples

We illustrate this estimation method using two example data sets, one from a two-color cDNA-array experiment and one from an experiment conducted using Affymetrix oligonucleotide arrays. The first example comes from a toxicology experiment by Bartosiewicz *et al.* (2000) in which male Swiss Webster mice were injected with a toxin. We shall focus on a single slide from this experiment. For this array, the treatment mouse was injected with 0.15 mg/kg of β-napthoflavone dissolved in 10 ml/kg of corn oil, and the control mouse was injected with 10 ml/kg of corn oil. mRNA from the livers of these mice was reverse transcribed and fluor labelled, with the treatment sample labelled with Cy5 and the control sample labelled with Cy3. The samples were hybridized to a spotted cDNA array on which each of the 138 genes was replicated between six and 14 times, resulting in a total of 1008 spots.

For the mouse data, we will model the differences of the transformed control and treatment observations rather than the transformed observations themselves. The difference of the transformed observations from replicate j of gene i, $\Delta h_{\lambda ij}$, can be modeled as

$$\Delta h_{\lambda ij} = \mu_i + \varepsilon_{ij}, \tag{8}$$

where μ_i is a gene effect and ε_{ij} is a normally distributed error term. Notice that (8) is a one-way ANOVA model. (Also note that in the case of differences, the sample size for the linear model will not be the total number of observations, but the total number of differences).

Figure 1 shows the partially maximized log likelihood for the mouse data as a function of the transformation parameter, λ. The likelihood is maximized at $\lambda = 1.13 \times 10^9$. Note that the median value of z is 4.6×10^4, which is on the order of $\sqrt{\hat{\lambda}}$, which is 3.7×10^4, suggesting that the squared median of the data might be a plausible starting value for of λ for Newton's method or other algorithms.

An asymptotic 95% confidence interval for the MLE, $\hat{\lambda}$, consists of those values of λ for which

$$l_{\max}(\hat{\lambda}) - l_{\max}(\lambda) < \frac{1}{2}\chi^2_{1,.05},$$

where $\chi^2_{1,.05}$ is the upper 5% quantile of a χ^2_1 distribution (Box and Cox (1964)). This yields a confidence interval for λ of $(8.67 \times 10^8, 1.47 \times 10^9)$.

Figure 2 shows a quantile–quantile plot of the residuals from the linear model (8) fit to the transformed data versus a standard normal distribution. The residuals appear to

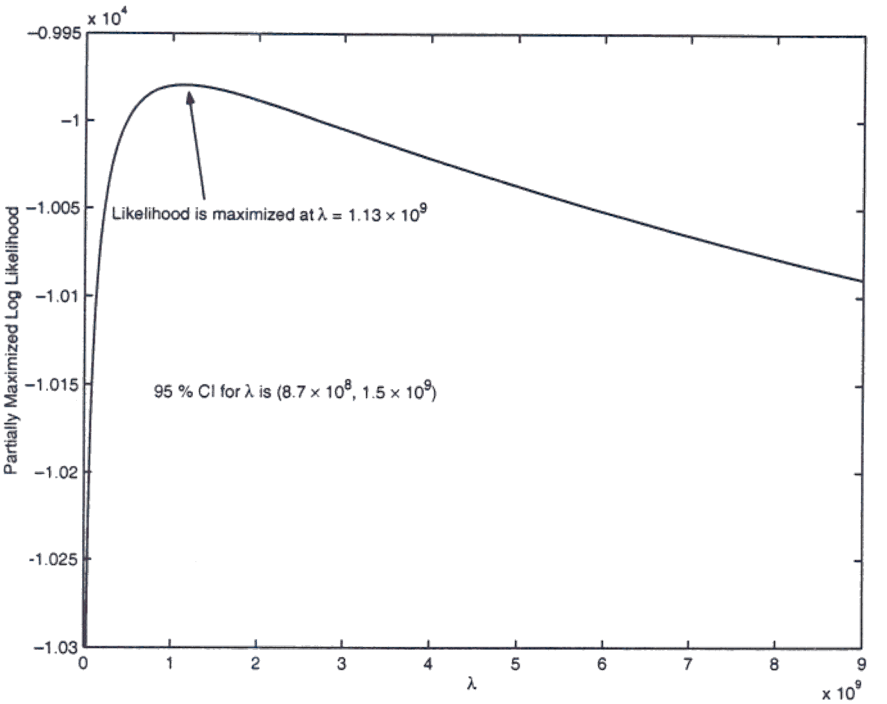

Fig. 1. Log likelihood by transformation parameter, mouse data.

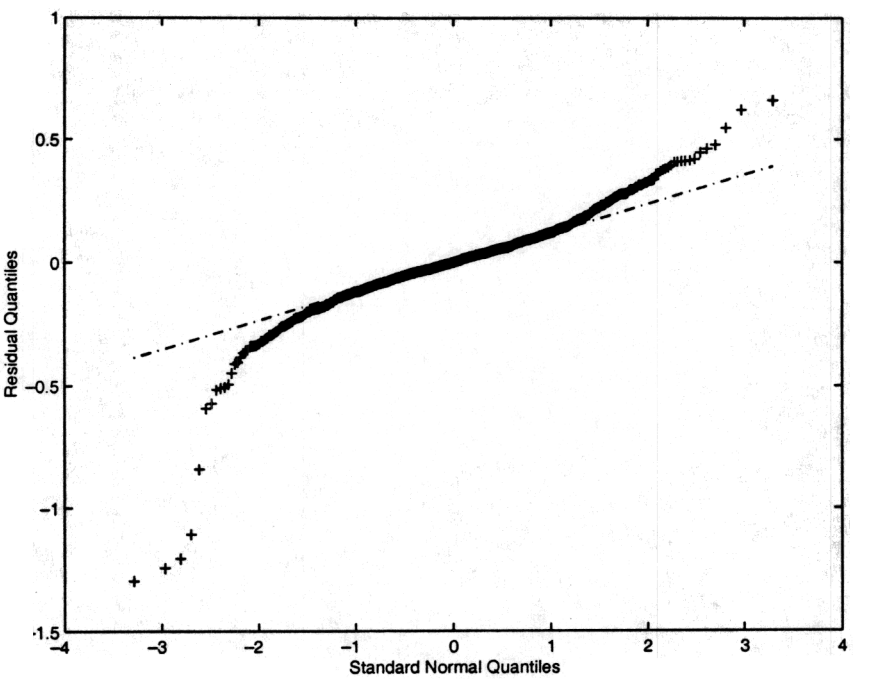

Fig. 2. QQ plot of residuals versus standard normal, maximum-likelihood transformation, mouse data.

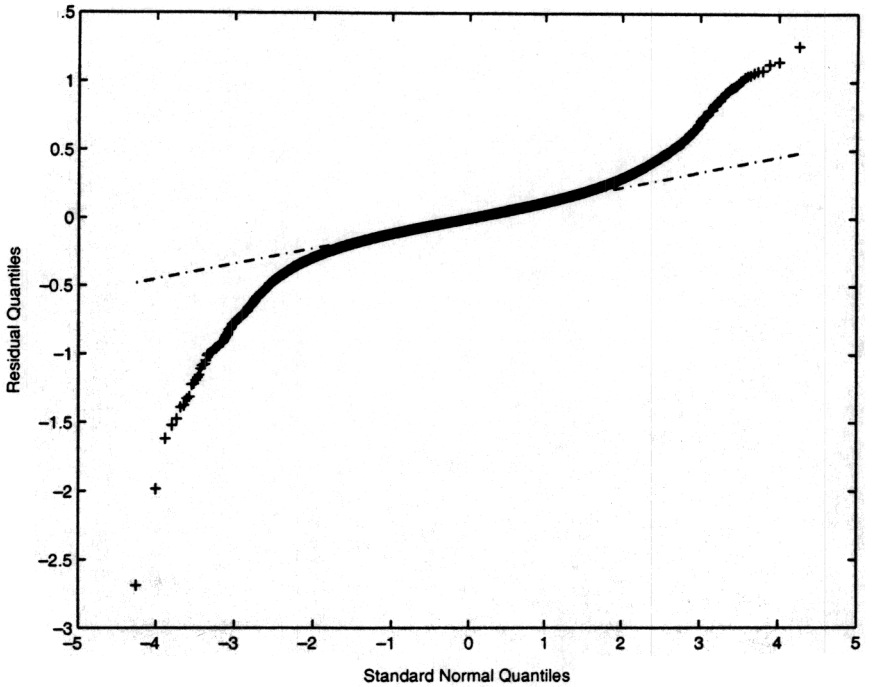

Fig. 3. QQ plot of residuals versus standard normal, maximum-likelihood transformation, autism data.

come from a distribution with heavier tails than a normal distribution. Although the plot appears to exhibit some skewness, this is entirely due to the four observations in the lower left-hand corner. These observations appear to be outliers resulting from phenomena such as dust on the slide, since they all occur in genes which are expressed near background in the control data, and feature a single observation which differs so hugely from the other replicates that it is unlikely to result from actual gene expression. (These observations will be excluded from the analysis of Section 3.) Examination of residuals from the linear model appears to facilitate identification of outlying observations, since these outliers were much more obvious in the residuals than they would have been in the raw data.

When the maximum-likelihood estimation is conducted again after removal of the 4 outliers, $\hat{\lambda} = 9.67 \times 10^8$, with a 95% confidence interval of $(7.67 \times 10^8, 1.22 \times 10^9)$.

The second example comes from an experiment using four Affymetrix HG_U95 arrays, which is described in Geller *et al.* (2003). In this experiment, a lymphoblastoid cell line from a single autistic child was grown up in four separate flasks. RNA extraction, cDNA synthesis, and in-vitro labelling were conducted separately on each of the four samples, and each of the samples was hybridized to a separate array.

For the autism data, we model the transformed perfect match minus mismatch observation from gene j on array i, $h_{\lambda ij}$, as

$$h_{\lambda ij} = \mu_i + \eta_j + \varepsilon_{ij}, \tag{9}$$

where μ_i is a fixed array effect, η_j is a fixed gene effect, and ε_{ij} is a normally distributed error term. Notice that our model is a two-factor ANOVA model without an interaction term. (We cannot fit the interaction term due to the absence of replicated genes, but we would not expect a gene-array interaction effect anyway.)

For the autism data, the likelihood is maximized at $\lambda = 3870$, and a 95% confidence interval for $\hat{\lambda}$ is (3750, 4000). For these data $\hat{\lambda}$ lies in between the squared median of the data, which is 900, and the squared mean, which is 32 400. Figure 3 shows a quantile–quantile plot of the residuals from the linear model (9). The residuals, again, appear to come from a symmetric distribution with tails heavier than a normal distribution.

3 OTHER METHODS OF ESTIMATING THE TRANSFORMATION PARAMETER

Maximum-likelihood estimation of the transformation parameter in the manner described above in essence simultaneously optimizes constancy of variance, the fit of the transformed residuals to a normal distribution, and the

fit to the linear model. In some applications, some of these criteria may be more important than others. For example, for many traditional statistical techniques data that are symmetric are almost as good as data that are normally distributed, and by trying to force the transformed data to fit all of the moments of a normal distribution we may inadvertently compromise those characteristics in which we are most interested. In such cases, we may search within the family of generalized-log transformations for a transformation optimizing the quantity of interest, simply by minimizing the appropriate statistic.

For example, to find a transformation minimizing the skewness of residuals from the linear model, we would look for a transformation for which the skewness coefficient of the residuals is equal to 0. To find a transformation for which the fixed effects in an ANOVA model are the most linear, we would look for a transformation minimizing the F-statistic for the interaction term in the model. (Notice that the two estimation procedures just mentioned both incorporate the linear model structure used in the maximum-likelihood estimation.) To find a transformation minimizing the dependency of the replicate mean on the replicate variance, we would regress the replicate standard deviation of the transformed data on the replicate mean and look for the transformation minimizing the absolute value of the t-statistic for the significance of the slope parameter. These other optimal transformations also provide a means of assessing the quality of the maximum-likelihood estimate of the transformation parameter. If the MLE differs too greatly from the optimal transformation parameter under another criterion, this might be cause for concern.

We illustrate the skewness-minimizing transformation and the transformation minimizing dependency of the replicate mean and variance using the mouse data. For these data, the skewness coefficient is non-monotonic in the transformation parameter, so there are two values of λ for which the skewness coefficient is equal to 0, which are 2.31×10^7 and 2.27×10^8. A asymptotic 95% confidence interval for the skewness-minimizing transformation consists of those values of λ for which the absolute skewness coefficient is not significant at the 5% level. For a sample of size 1004 the absolute skewness is not statistically significant if it is less than 0.1515, which yields the confidence interval $(2.71 \times 10^6, 2.00 \times 10^9)$. This interval includes the maximum likelihood transformation, ($\lambda = 9.67 \times 10^8$ following removal of the outliers), implying the the MLE does provide sufficient symmetry.

A transformation minimizing mean-variance dependency may be found by minimizing the t-statistic of the regression of the replicate standard deviation on the replicate mean. For the mouse data, the t-statistic is equal to 0 at $\lambda = 4.03 \times 10^9$. An asymptotic 95% confidence interval for the t-minimizing transformation consists of those values of λ for which the t-statistic is not significant at the 5% level. For 136 degrees of freedom (since we have 138 genes and lose 2 degrees of freedom from fitting the regression parameters) the cutoff for significance of the t-statistic is ± 1.9776, which yields the confidence interval $(2.02 \times 10^9, 8.09 \times 10^9)$. This confidence interval excludes the maximum-likelihood transformation, where $\hat{\lambda} = 9.67 \times 10^8$.

However, examination of Figure 4, showing the replicate mean and standard deviation for the maximum-likelihood transformation and the t-minimizing transformation (in the lower left-hand and upper-right-hand panels, respectively) indicates that both of these transformations provide reasonable variance stabilization. For comparison, the lower-right-hand panel shows log ratios of the same data, which shows dramatic mean-variance dependency.

It is perhaps surprising that the maximum-likelihood transformation provides good variance-stabilization and symmetrization in light of the fact that the normal likelihood is almost certainly the 'wrong' likelihood for the transformed data, given the apparent heavy-tailed distribution of the transformed residuals. However, the data appear (anecdotally speaking) to be somewhat insensitive to which member of the generalized-log family is used.

Robust statistical methods could certainly be used to address the heavy-tailedness of the residual distribution. However, robust methodologies tend to be much more computationally expensive and hence slower to run than the methods presented here. Given that the normal-theory approach gives reasonable results, the additional effort of using robust methods may not be worthwhile for the researcher interested in obtaining a quick answer.

4 CONCLUSIONS

The generalized-log transformation of Durbin *et al.* (2002), Huber *et al.* (2002) and Munson (2001) with parameter $\lambda = a^2/b^2$ stabilizes the variance of data where $\mathrm{Var}(z) = a^2 + b^2E^2(z)$. Maximum-likelihood estimation in the manner of Box and Cox (1964) can be used to estimate a transformation parameter for data where observations have different values of a and b. This procedure estimates the transformation parameter while simultaneously fitting a linear model to the data, allowing for easy and quick estimation of the transformation parameter (by minimizing the error sum of squares of the linear model fit to the transformed data) while accounting for the experimental structure of the data. Transformations minimizing residual skewness, mean-variance dependency, and other criteria may be found by minimizing the appropriate statistic. The maximum likelihood estimate appears to perform well compared to

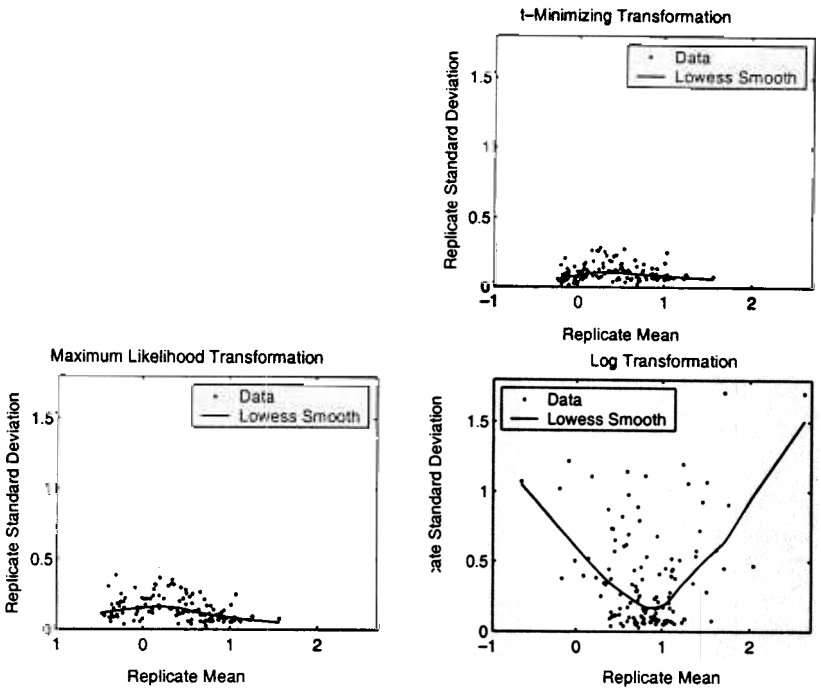

Fig. 4. Replicate mean and standard deviation of differences of transformed observations, three different transformations.

transformations specifically minimizing residual skewness and mean-variance dependency, especially in light of the fact that the normal likelihood is a first approximation to the 'true' distribution of the transformed data and was chosen primarily for computational convenience.

ACKNOWLEDGEMENTS

The research reported in this paper was supported by grants from the National Science Foundation (ACI 96-19020, and DMS 98-70172) and the National Institute of Environmental Health Sciences, National Institutes of Health (P43 ES04699).

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. *J. R. Stat. Soc.*, Series B *(Methodological)*, **26**, 211–252.

Durbin,B., Hardin,J., Hawkins,D.M. and Rocke,D.M. (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18**, 105S–110S.

Ferguson,T.S. (1996) *A Course in Large Sample Theory*. Chapman and Hall, London.

Geller,S.C., Gregg,J.P., Hagermann,P. and Rocke,D.M. (2003) Transformation and normalization of oligonucleotide microarray data. *Bioinformatics*, **19**, in press.

Huber,W., von Heydebreck,A., Sültmann,H., Poustka,A. and Vingron,M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, 96S–104S.

Kerr,K., Martin,M. and Churchill,G. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

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 of Low Level Analysis of Affymetrix GeneChip Data*.

Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol.*, **8**, 557–569.

APPENDIX: NUMERICAL OPTIMIZATION VIA NEWTON'S METHOD

Newton's method provides a means of finding a root of a smooth function. We use this technique to perform numerical minimization of the error sum of squares by finding a root of the first derivative of SSE(λ). Plots of the likelihood may be used to confirm that the root found does

indeed constitute a global maximum. Denote $\frac{\partial}{\partial\lambda}\mathrm{SSE}(\lambda)$ by $\mathrm{SSE}'(\lambda)$ and $\frac{\partial^2}{\partial\lambda^2}\mathrm{SSE}(\lambda)$ by $\mathrm{SSE}''(\lambda)$. A new estimate of λ, $\lambda_{(n+1)}$, may be obtained from the previous estimate, $\lambda_{(n)}$ using

$$\lambda_{(n+1)} = \lambda_{(n)} - \frac{\mathrm{SSE}'(\lambda_{(n)})}{\mathrm{SSE}''(\lambda_{(n)})}. \tag{10}$$

Convergence is achieved when $|\mathrm{SSE}'(\lambda)|$ is less than the predetermined application tolerance.

For the generalized log transformation with parameter λ,

$$\mathrm{SSE}'(\lambda) = 2\sum_{i=1}^{n}(w_{\lambda i} - \hat{w}_{\lambda i})(w'_{\lambda i} - \hat{w}'_{\lambda i}), \tag{11}$$

where

$$\hat{w}_{\lambda i} = \boldsymbol{x}_i^{\top}\hat{\beta}(\lambda),$$

$\boldsymbol{x}_i^{\top}$ is the ith row of the design matrix,

$$\begin{aligned} w'_{\lambda i} &= \frac{\partial}{\partial\lambda}w_{\lambda i} \\ &= [2\{z_i + \sqrt{z_i^2+\lambda}\}\sqrt{z_i^2+\lambda}]^{-1}J^{-1/n}(\lambda) \\ &\quad + \ln(z_i + \sqrt{z_i^2+\lambda})\frac{\partial}{\partial\lambda}J^{-1/n}(\lambda), \end{aligned}$$

where

$$\frac{\partial}{\partial\lambda}J^{-1/n}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}J^{-1/n}(\lambda)/\{2(z_i^2+\lambda)\},$$

and

$$\begin{aligned} \hat{w}'_{\lambda i} &= \frac{\partial}{\partial\lambda}\hat{w}_{\lambda i} \\ &= \boldsymbol{x}_i^{\top}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{w}'_{\lambda}, \end{aligned}$$

The second derivative of the error sum of squares is

$$\begin{aligned} \mathrm{SSE}''(\lambda) &= 2\sum_{i=1}^{n}(w'_{\lambda i} - \hat{w}'_{\lambda i})^2 \\ &\quad + 2\sum_{i=1}^{n}(w_{\lambda i} - \hat{w}_{\lambda i})(w''_{\lambda i} - \hat{w}''_{\lambda i}), \end{aligned}$$

where

$$\begin{aligned} w''_{\lambda i} &= \frac{\partial^2}{\partial\lambda^2}w_{\lambda i} \\ &= -\frac{1}{4}J^{-1/n}(\lambda)(z_i + \sqrt{z_i^2+\lambda})^{-1}\{z_i^2+\lambda\}^{-\frac{3}{2}} \\ &\quad -\frac{1}{4}J^{-1/n}(\lambda)(z_i + \sqrt{z_i^2+\lambda})^{-2}\{z_i^2+\lambda\}^{-1} \\ &\quad +(z_i + \sqrt{z_i^2+\lambda})^{-1}\{z_i^2+\lambda\}^{-\frac{1}{2}}\frac{\partial}{\partial\lambda}J^{-1/n} \\ &\quad + \ln(z_i + \sqrt{z_i^2+\lambda})\frac{\partial^2}{\partial\lambda^2}J^{-1/n} \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^2}{\partial\lambda^2}J^{-1/n}(\lambda) &= \frac{1}{2n}\sum_{i=1}^{n}\{z_i^2+\lambda\}^{-1}\frac{\partial}{\partial\lambda}J^{-1/n}(\lambda) \\ &\quad -\frac{1}{2n}\sum_{i=1}^{n}\{z_i^2+\lambda\}^{-2}J^{-1/n}(\lambda) \end{aligned}$$

and

$$\begin{aligned} \hat{w}''_{\lambda i} &= \frac{\partial^2}{\partial\lambda^2}w_{\lambda i} \\ &= \boldsymbol{x}_i^{\top}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{w}''_{\lambda}. \end{aligned}$$