Likelihood Based Evaluation of Normalization Methods

Bonnie LaFleur Dean Billheimer Heidi Chen

May 26, 2005

Abstract

A suite of emerging measurement technologies in biology, and elsewhere, provide unprecedented ability to measure individual components of complex systems. DNA microarrays, mass spectrometry, raman spectrometry, and gel electrophoresis techniques, share the characteristics of making highly multivariate measurements, but are limited in that only "relative intensity" is measured (typically in arbitrary units). Further, nuisance variation affecting an entire observation unit (e.g., array or spectrum) is frequently present, and complicates data anlaysis. Although many normalization methods have been proposed to deal with these problems, evaluation of a method, and choosing between competing methods remain open problems. We have developed an approach to evaluate normalization methods for high-throughput measurement technologies. Our approach is statistically based, and can be viewed as an extension of widely used Box-Cox analysis of transformations. We describe our likelihood based evaluation metric, and illustrate its use for several recently proposed cDNA microarray normalization methods. Our proposed metric corresponds strongly with normalization diagnostic plots in cases where the plots indicate a preferred method. More importantly, the likelihood based metric provides guidance in situations in which diagnostic plots are ambiguous or unavailable.

1 Introduction

Many measurement techniques used in biomedical research, including DNA–based microarray analysis, gel electrophoresis, chromatography, and mass and Raman spectroscopy share the characteristics of making highly multivariate measurements, but are limited to measuring only the *relative intensity* of each response. Frequently, non-biologic sources of variation also affect the measured outcomes, and are present at the level of the multivariate observation (e.g., at each microarray chip, gel, or spectrum). Such "nuisance" variation includes non-specific background or baseline, intensity scaling, and nonlinear response, among others. When combined, the arbitrary intensity scale and nuisance variation limit our ability to conduct quantitative data analysis.

To address this problem, normalization methods have been applied with the goal of reducing the effects of nuisance variation, and making response intensities comparable across (multivariate) observations. Normalization is a mathematical transformation of the measured intensities, and may include steps such as background or baseline subtraction, smoothing, multiplicative scaling and possibly other nonlinear transformations. Despite the wide variety of normalization methods that have been proposed, surprisingly little information is available regarding how one should select a normalization method, or should choose between competing normalizations. At best, there appear to be diagnostic plots that indicate (heuristically) whether the applied normalization is adequate.

To focus the discussion, we address normalization methods proposed for DNA–based microarray experiments. Normalization methods for these experiments are well documented for both cDNA and oligonucleotide arrays. Methods specific to cDNA experiments can be found in Yang *et al.* (2002) and further discussed in Cui *et al.* (2003). Techniques for oligonucleotide

arrays are also available in Schadt *et al.*(2000), Bolstad *et al.* (2003) and are similar in principle to those used for cDNA or two-color experiments. Further, software for performing these transformations is widely available for both cDNA and oligonucleotide microarray (MAS 5.0, the algorithm incorporated into the Affymetrix system; the dChip software of Li and Wong (2001); and the microarray software available through the Bioconductor project (Gentleman *et al.* (2003), for example). Quackenbush (2002) provides a nice review of general methods of normalization for both cDNA and oligonucleotide data.

In the microarray setting, the ability to formally assess and evaluate competing normalization techniques is lacking. Most researchers use plots to compare normalization techniques, the most common being similar to the Bland and Altman plots (Bland and Altman, 1986). These are termed RI plots (for ratio versus intensity), or MVA plots (log intensity ratio versus abundance). For cDNA arrays, the typical format is to plot the log_2 ratio of red and green dye channels ($\log_2 R/G$) against the $\log_2 \sqrt{(RG)}$. Similarly, for oligonucleotide arrays, one usually plots pairwise differences (two chips) versus the pairwise averages. These plots have proven to be an effective method for detection of systematic relationships between pair differences and pair averages (on the $\log_2$ scale), background differences, nonlinearity between data pairs, or even spatial heterogeneity. Under the assumption that there are no expression differences for most genes, this graph shows a linear band centered horizontally around zero, denoting the (random) differences between dyes or pairs.

In this paper we present and illustrate the use of a normalization scoring criterion based on the Gaussian likelihood and probability densities induced from data transformations. We consider this criterion is a simple extension of the analysis of transformations introduced by Box and Cox (1964). Any novelty derives from noticing that normalization is a special form of transformation. We demonstrate that the proposed criterion conforms to our heuristic judgment

of "good" normalization methods, and equally importantly, provides a quantitative assessment for comparing competing methods.

Our goal is to lay the foundation for a theory of normalization which uses standard statistical notation and solutions. Specifically, we provide initial steps to characterize the normalization problem, identify important technical issues, and provide an interpretable statistical framework. We first describe the relationship between the normalization problem and probability theory of data transformation. We then use simulation-based examples of cDNA microarray expression to show how this framework is applied to data. Recently proposed normalization methods are applied to these simulated data. RI plots, along with our likelihood based criterion, are presented to examine the effect of various transformations on removing nuisance variation. Although we focus on normalization methods applied to microarray data, we note that normalization operations are used widely throughout science. The principles described here are broadly applicable.

2 Characterizing the Normalization Problem

There are a number of well established statistical techniques with a goal similar to that of normalization: removing nuisance variation. For example, analysis of covariance is a linear normalization method where response data are made more comparable by linear adjustment based on the magnitude of a measured covariate. Similarly, blocking in experimental design and stratification in sampling are techniques designed to remove variation via planned data collection. Here we outline the statistical characteristics of normalization problem. For illustration, it is convenient to use a simple additive model with multiplicative nuisance variation. We demonstrate that *a normalization procedure amounts to constraining a statistical model to achieve model identification.* Further, the choice of identifying constraint is strongly associated

with the choice of normalization method.

Consider the usual additive statistical model written as

$$\mathbf{x}_i = \boldsymbol{\theta} + \boldsymbol{\epsilon}_i. \tag{1}$$

where $\mathbf{x}_i$ is a p-vector of "ideal" observations, $\boldsymbol{\theta}$ is the p-vector mean of $\mathbf{x}_i$, $\boldsymbol{\epsilon}_i$ is additive error, and $i \in \{1, 2, \ldots, n\}$ indexes observation number.

Now, instead of $\mathbf{x}$, we observe a "corrupted" version that has been subjected to multiplicative nuisance variation.

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{bmatrix} = \begin{bmatrix} \alpha_1 & & 0 \\ & \alpha_2 & \\ 0 & & \ddots \\ & & & \alpha_n \end{bmatrix} \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} = \begin{bmatrix} \alpha_1 & & 0 \\ & \alpha_2 & \\ 0 & & \ddots \\ & & & \alpha_n \end{bmatrix} \left(\begin{bmatrix} \boldsymbol{\theta}' \\ \boldsymbol{\theta}' \\ \vdots \\ \boldsymbol{\theta}' \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_n \end{bmatrix} \right) \tag{2}$$

Written more compactly,

$$\mathbf{Y} = \boldsymbol{\alpha}\,\boldsymbol{\Theta} + \boldsymbol{\eta}$$

where $\boldsymbol{\alpha}$ is a diagonal matrix of n nuisance parameters and $\boldsymbol{\eta}$ is an $n \times p$ error matrix. In this example the nuisance parameters represent intensity scaling variation. However, this model is easily generalized to include baseline/background variation, mean-variance dependence, or non-Gaussian error distributions.

The model in equation (2) is clearly not identified since $\boldsymbol{\alpha}$ can be replaced with $\boldsymbol{\alpha}/c$, and $\boldsymbol{\Theta}$ with $c\,\boldsymbol{\Theta}$ for any value of $c \neq 0$ (Kadane, 1978). We need a model constraint to identify equation (2). This constraint simultaneously "selects" a normalization method. In this setting a typical heuristic normalization choice is to set the following:

$$\hat{\alpha_i} = \sum_{j=1}^{p} y_{ij} \tag{3}$$

Then, the associated normalization (transformation) to remove multiplicative nuisance variation becomes

$$\mathbf{z}_i = \frac{1}{\hat{\alpha}_i}\,\mathbf{y}_i \tag{4}$$

where $\mathbf{z}_i$ is the resulting normalized variable. The constraint used to identify the model can be written as, $\sum_{j=1}^{p} z_{ij} = 1$. Other choices of normalization criteria lead to different model identification constraints.

To choose between competing normalization methods, we refer to the seminal paper by Box and Cox (1964). Although the popular legacy of their paper is the well known Box-Cox transformation, a fundamental idea is that we can use the maximized *likelihood of the original data* to judge the quality of competing normalizations. This likelihood is easily calculated by computing the density of the transformed data and multiplying by the Jacobian of the transformation. Frequently our method of analysis "post-normalization" will be based on the Gaussian distribution. Therefore, the likelihood of the original data is given by the Gaussian density of the transformed data times the Jacobian of the transformation:

$$f(\mathbf{Y} \mid \boldsymbol{\alpha}, \hat{\boldsymbol{\theta}}_{\boldsymbol{z}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}) = \mid 2\pi\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} \mid^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{z}_i - \hat{\boldsymbol{\theta}}_{\boldsymbol{z}}\right)' \hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}^{\ -}\left(\mathbf{z}_i - \hat{\boldsymbol{\theta}}_{\boldsymbol{z}}\right)\right\} \mathbf{J}(\boldsymbol{\alpha}, \mathbf{y}) \tag{5}$$

where $\hat{\boldsymbol{\theta}}_{\boldsymbol{z}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}$ are the maximum likelihood estimates for the mean vector and variance-covariance matrix of the transformed data, respectively, and $\mathbf{J}(\boldsymbol{\alpha}, \mathbf{y})$ is the Jacobian of the transformation

$$\mathbf{J}(\boldsymbol{\alpha}, \mathbf{y}) = \text{abs}\ \left|\left(\frac{\partial \mathbf{z}_i}{\partial \mathbf{y}_j}\right)\right|$$

The constraints used in model identification induce a singularity in the variance-covariance matrix of the likelihood. Following Mardia *et al.* (1979) we use a generalized inverse in the likelihood to solve this technical difficulty. Finally, we note that different normalization methods may impose different numbers of constraints on the statistical model. An adjustment to the

likelihood, such as AIC or BIC (Schwarz, 1978) or other method accounting for the differing number of free parameters may be needed for comparing normalizations of different dimension. We explore this point further in the Discussion section.

3 Model-based cDNA Microarray Data Simulation

We demonstrate calculation of the likelihood-based scoring criterion using simulated cDNA microarray data. In this setting Cui *et al.*(2003) describe a general model for signal intensity with multiple error sources. We take their model and the transformations they examine as a starting point to demonstrate and evaluate our likelihood based normalization criterion. The transformations are intended to correct systematic variation in log ratios. The effectiveness of each transformation can be evaluated by means of an RI plot. Our goal is to show that the likelihood based scoring criterion matches our heuristic interpretation of a "good" transformation. In addition, we note that the scoring criterion easily extends to situations in which diagnostic plots are difficult to interpret or to construct.

The basic premise behind cDNA array analysis is the examination of the relationship between biologic samples of messenger RNA (mRNA). The cDNA (cloned copies of mRNA) is isolated from specimens of various disease states, subjects, etc. and then labeled and hybridized to arrays (slides) that have known DNA sequences (called spots or genes) immobilized on them. The cDNA samples are often called "targets" and the DNA that the sample is hybridized to are called "probes". When there is only a single specimen, one dye used to measure the fluorescent intensity of the binding; for two types of specimens (e.g., disease versus normal) then typically two dyes are used on a single slide. The relative intensity of the fluourescent dye(s) (or channels) are measured, and these intensities are thought to describe the mRNA concentrations from the

cDNA isolates. Generally, the ratio of the two fluorescent signals at each spot (gene) is used to describe the increase (up regulation) or decrease (down regulation) of gene expression in a specimen. Cui *et al.*(2003) describe the ideal cDNA experiment, where Y_{ik} is the observed fluorescence intensity detected from both $i = r$ or g channels and $k = 1, \ldots, K$ spots. That is,

$$Y_{ik} = \alpha_i + \beta_i X_{ik} \tag{6}$$

Where the signal at channel i and gene k is comprised of the background signal, α_i, the concentration of the signal itensity, X_{ik}, and the slope of the linear relationship, β_i. Deviations from this model may include multiplicative and/or additive errors, and the data are more realistically described by the models specified in Cui *et al.* (2003) and Rocke and Durbin (2001):

$$Y_{ik} = \alpha_i + \beta_i X_{ik} e^{\eta_k + \zeta_{ik}} + \epsilon_k + \delta_{ik}. \tag{7}$$

As in Cui *et al.* (2003), we base our simulations on (7), where X_{ik} is drawn from a lognormal distribution with mean 7 and standard deviation 1.1, the multiplicative errors, η_k and ζ_{ik}, are drawn from independent normal distributions, $N(0,\, \sigma_\eta^2)$ and $N(0,\, \sigma_{\zeta_i}^2)$, respectively. The additive errors, ϵ_k and δ_{ik}, were drawn from normal distributions, $N(0,\, \epsilon_\eta^2)$ and $N(0,\, \sigma_{\delta_i}^2)$, respectively. Using this model formulation, we can vary background differences, error structures (both additive and multiplicative), and slope (intensity) differences by modifying simulation parameters. We explore data simulations where either $\alpha_g \neq \alpha_r$ or $\beta_g \neq \beta_r$, where α_i denotes background signal, and β_i denotes the channel slope. We do not vary the error component parameters to introduce distortion based on multiplicative or additive error (these simulation parameters are held constant). Graphical examination of the effect of transformation (using RI plots) is then compared with our quantitative assessment (using the likelihood).

4 Data Transformations and Calculations of the Jacobians

In this section we describe the transformations to be compared, and demonstrate calculation of the Jacobians required in (5). For background, appropriateness and justifications for each of the transformations below, we refer interested readers to the original papers. We use these transformations solely for illustrative purposes and are not advocating any particular transformation for general use.

4.1 Shift Transformation

The shift transformation, also called the shift-log, is described in Kerr *et al.*(2002). For each gene, a constant is added to the log expression in one channel, and the same constant is subtracted from the other channel:

$$\begin{aligned} Z_{rk} &= \log_2(Y_{rk} - C) \\ Z_{gk} &= \log_2(Y_{gk} + C) \end{aligned} \tag{8}$$

The value of the constant, C, is calculated as follows:

$$C = \overset{argmin}{C} \sum_{k=1}^{K} \left| \log_2\left(\frac{Y_{rk} - C}{Y_{gk} + C} \right) - \log_2 \left(\frac{\widetilde{Y_{rk}}}{Y_{gk}} \right) \right| \tag{9}$$

where the tilde denotes the median log ratio of the array, and K is the total number of genes on the array.

To construct the Jacobian for this transformation we first require the following partial derivatives.

$$\frac{\partial Z_{rk}}{\partial Y_{rk}} = \frac{1}{(Y_{rk} - C)\log(2)}$$

$$\frac{\partial Z_{gk}}{\partial Y_{gk}} = \frac{1}{(Y_{gk}+C)\log(2)} \tag{10}$$

Finally, we compute the Jacobian as the product of absolute values of partial derivatives. (This same form for the Jacobian is used for all ensuing transformations.)

$$J = \prod_{k=1}^{K}\prod_{i=r}^{g}\left|\frac{\partial Z_{ik}}{\partial Y_{ik}}\right| \tag{11}$$

4.2 Curve Fitting Transformations

Curve fitting transformations contrain the local mean of intensity log ratios (in an RI plot) to zero. A linear or nonlinear function is fit to the log ratios, and the fitted value for each gene is subtracted from the expression log ratio. A commonly used curve fitting technique uses locally weighted linear regression (lowess) and is detailed in Yang *et al.* (2002)yang:dudoit. The transformation is defined as

$$\begin{aligned} Z_{rk} &= \log_2(Y_{rk}) + C_k/2 \\ Z_{gk} &= \log_2(Y_{gk}) - C_k/2 \end{aligned} \tag{12}$$

Where the constant, C_k, is a gene-specific constant obtained from the fitted curve.

To compute the Jacobian we assume C_k is constant, and write the partial derivative as follows:

$$\frac{\partial Z_{ik}}{\partial Y_{ik}} = \frac{1}{Y_{ik}\log(2)} \tag{13}$$

4.3 Arsinh Transformation

The arsinh transformation is a variance-stabilizing transformation described in Huber *et al.* (2003). It is intended to de-couple the mean-variance dependence often observed in microarray

expression data. For replicate cDNA analyses in particular, the variance of the spot intensities increases with the mean (Rocke and Durbin, 2001). A funnel shape in the RI plot is indicative of such a mean-variance relationship.

The channel specific transformation is

$$Z_{ik} = \log(\beta_i Y_{ik} + C_i + \sqrt{(\beta_i Y_{ik} + C_i)^2 + 1}) \tag{14}$$

where $i = r$ or g for red and green channels. The parameters b_i and C_i are estimated by means of a robust variant of likelihood maximization Huber *et al.* (2002).

This transformation has the following Jacobian:

$$\frac{\partial Z_{ik}}{\partial Y_{ik}} = \frac{\beta_i}{\beta_i Y_{ik} + C_i + \sqrt{(b_i Y_{ik} + C_i)^2 + 1}}(1 + \frac{\beta_i Y_{ik} + C_i}{\sqrt{(\beta_i Y_{ik} + C_i)^2 + 1}}) \tag{15}$$

The goal of this transformation is to achieve constant coefficient of variation. For high intensity values the transformation approximates logarithmic transformation, and for low intensity values the transformed values it tends to zero. This transformation also corrects for curvature due to background or slope differences.

4.4 Linlog Transformation

The linlog transformation combines the logrithmic transformation with a linear transformation, depending on channel specific spot intensity. For high intensity spots the data are transformed using the logrithmic transformation, and linear transformation is used for low intensity spots. The tranformation function is

$$Z_{ik} = \text{linlog}(Y_{ik}) = \begin{cases} \frac{\log_2(d_i)-1}{\ln 2} + \frac{Y_{ik}}{(d_i \times \ln 2)} & Y_{ik} < d_i \\ \log_2(Y_{ik}) & Y_{ik} \geq d_i \end{cases} \tag{16}$$

Again, this is channel specific and $i = r$ or g. The intensity, d_i is estimated by minimizing the absolute deviation of the IQR of log ratios from a linear range of values the median IQR of the

entire array. Cui *et al.* (2003) choose bins for the linear range that place 25-30% of the data in a range.

The resulting partial derivatives are

$$\frac{\partial Z_{ik}}{\partial Y_{ik}} = \begin{cases} \frac{1}{(d_i \times \ln 2)} & Y_{ik} < d_i \\ \frac{1}{(y_{ik} \times \ln 2)} & Y_{ik} \geq d_i \end{cases}$$

This transformation does not correct for curvature due to background or slope differences, so a combination between the linlog and the shift transformation, named linlogshift is proposed by Cui *et al.* (2003) this transformation function has the following form:

$$\begin{aligned} Z_{rk} &= \text{linlog}(Y_{rk} - C) \\ Z_{gk} &= \text{linlog}(Y_{gk} + C) \end{aligned} \tag{17}$$

and the resulting partial derivatives are:

$$\frac{dZ_{rk}}{dY_{rk}} = \begin{cases} \frac{1}{(d_r \times ln2)}) & Y_{rk} - C < d_r \\ \frac{1}{((y_{rk} - C) \times ln2)}) & Y_{rk} - C \geq d_r \end{cases}$$

$$\frac{dZ_{gk}}{dY_{gk}} = \begin{cases} \frac{1}{(d_g \times ln2)}) & Y_{gk} + C < d_g \\ \frac{1}{((y_{gk} + C) \times ln2)}) & Y_{gk} + C \geq d_g \end{cases} \tag{18}$$

5 Examples and Results

Data were simulated using the model described by (7) with X_k randomly drawn from a lognormal distribution with mean of 7 and standard deviation 1.1. The multiplicative errors, η_k and ζ_{ik}, were drawn from normal distributions, $N(0, \sigma_\eta^2 = 0.01)$ and $N(0, \sigma_{\zeta_i}^2 = 0.01)$. The additive errors, ϵ_k and δ_{ik}, were drawn from normal distributions, $N(0, \epsilon_\eta^2 = 100)$ and $N(0, \sigma_{\delta_i}^2 = 100)$. Four

simulation models were examined, two with no replication (one array, one specimen) but with either background difference or slope difference, and two with background or slope difference and two replicate arrays (two arrays, one specimen). For each of the four models we examine the RI plot for the raw data, and each of the transformations discussed above. Additionally, we present the log-likelihood for the original (raw) data, the log-likelihood for the transformed data, both AIC and BIC, and the calculated value of the log-Jacobian (to reveal the contribution of each component). We include the AIC and BIC values to penalize more complicated transformations involving large numbers of parameters (e.g., with the lowess transformation). In effect, more fitted parameters in a transformation model result in a greater number of constraints in the transformed data. The likelihood-based values correspond to our qualitative evaluation of the graphical representation, and are more suitable when the graphical results are ambiguous.

Background Differences with One Array

In this model, we let $\alpha_r = 80$, $\alpha_g = 150$, and $\beta_r = \beta_g = 1$. The introduction of a difference in background between the red and green channels induces slight curvature as demostrated in figure 1. All of the transformations, with the exception of linlog, decrease curvature dramatically. That linlog does not perform as well is not surprising since its primary use is not in curvature reduction. It is difficult to determine which of the other transformations is "better" by examination of the RI plots. Table 1 shows the log-likelihood of the original data for the these transformations, and the smallest log-likelihood corresponds to the lowess transformation. Interestingly, BIC shows that the shift transformation actually has the lowest value after adjusting for the number of parameters used in the transformation.

Table 1: One Array with Background Differences

Transformation	LL for Y[a]	AIC	BIC	LL for Z[b]	LogJacobian
Shift	−66732	−66733	**−66736**	1083	−67815
Lowess	**−66708**	**−66725**	−66786	1091	−67799
LinLog	−69572	−69574	−69582	−233	−69339
LinLogShift	−67552	−67555	−67566	1791	−69344
Arsinh	−67976	−67980	−67994	2272	−70248

[a]Log-likelihood for the original data

[b]Log-likelihood for the transformed data

Slope Differences with One Array

In this model, we let $\beta_r = 0.05$, $\beta_g = 1$, and $\alpha_r = \alpha_g = 80$. The introduction of a slope difference for the red and green channels induces marked curvature, as demostrated in figure 2. Qualitative assessment of the transformations used for these data is slightly more useful, with lowess and arsinh outperforming the other methods. Again, it is difficult to determine whether the lowess or arsinh transformation is preferable by visual inspection alone. The likelihood methods, shown in Table 2, show the lowess transformation has the smallest log-likelihood value.

Examples with Replicate Arrays

For the remaining two examples, we use the same simulation parameters as the previous two examples. The difference now is that we observe replicate arrays on the same samples. Not suprisingly, with twice as many data points, the RI plots shown in figures 3 and 4 are more dense, making it diffult to evaluate subtle differences between the various transformation methods. The likelihood-based evaluations, shown in Tables 3 and 4 are effectively the same as the previous

Table 2: One Array with Slope Difference

Transformation	Yllike[a]	AIC	BIC	Zllike[b]	LogJacobian
Shift	−72288	−72289	−72293	−17236	−55052
Lowess	**−57556**	**−57573**	**−57635**	−572	−56984
LinLog	−74271	−74273	−74280	−16186	−58084
LinLogShift	−73631	−73634	−73645	−17282	−56349
Arsinh	−62539	−62543	−62558	3106	−65646

[a]Log-likelihood for the original data

[b]Log-likelihood for the transformed data

examples. A minor exception occurs for the AIC and BIC assessments. An overall increase in sample size results in the shift transformation having the greatest AIC and BIC values.

Table 3: Two Arrays with Background Difference

Transformation	Yllike[a]	AIC	BIC	Zllike[b]	LogJacobian
Shift	−120469	**−120470**	**−120474**	15035	−135504
Lowess	**−120461**	−120478	−120544	15010	−135471
LinLog	−125068	−125070	−125078	13665	−138734
LinLogShift	−120497	−120500	−120512	18245	−138743
Arsinh	−120470	−120474	−120489	19610	−140080

[a]Log-likelihood for the original data

[b]Log-likelihood for the transformed data

Table 4: Two Arrays with Slope Difference

Transformation	Yllike[a]	AIC	BIC	Zllike[b]	LogJacobian
Shift	−150652	−150653	−150657	−39659	−110992
Lowess	**−102281**	**−102298**	**−102365**	11592	−113873
LinLog	−152657	−152659	−152667	−36461	−116196
LinLogShift	−153489	−153492	−153504	−39928	−113561
Arsinh	−104778	−104782	−104798	26624	−131403

[a]Log-likelihood for the original data

[b]Log-likelihood for the transformed data

6 Discussion

We have formalized the method of normalization based on statistical principles that allow for quantitative assessment of current data transformations used in scientific measurement techniques. Having an estimate of quality of transformation can make selection among competing normalization methods more scientific and defensible. Currently, these normalization methods are chosen using qualitative methods, such as RI plots, which may result in ambiguous differences upon visual inspection.

Normalization methods commonly used in biology can be viewed as data transformations to meet prespecified constraints. These constraints are imposed to identify a statistical model in the presence of nuisance variation. Once viewed in this manner, a statistical modeling approach can be implemented to evaluate normalization. A simple approach is to view normalization as a particular kind of transformation, and use the resulting log-likelihoods as a metric to assess the quality of the transformation. We have extended the methods outlined in Box and Cox (1964) for multivariate responses, which allows us to calculate a log-likelihood for the original

data (based on the transformation).

Some normalization transformations require fitting a large number of parameters (e.g., lowess transformation), and subsequently induce a large number of constraints. Heuristically, it is reasonable to adjust the resulting log likelihood for the number of parameters used in fitting the transformation. We have elected to use AIC and BIC as two possible criteria for penalizing the number of parameters. However, we believe that parameter estimation for normalization is somewhat different than the long-standing model selection problem. Hence, AIC and BIC may not be optimal choices for the normalization problem. This topic is clearly an area of future research.

We examine some of the most common transformations used in cDNA or two-dye microarray experiments mainly because these experiments are easily presented in a model-based form, and the transformations are recognized by many researchers. However, our primary aim in this paper is to outline a theory of normalization that can be applied to many measurement techniques and transformations. We believe that by formulating normalization in a model-based manner, common statistical techniques (e.g., calculation of the log-likelihood) can, and should, be used to evaluate the effect of transformations on the data. These evaluations can be used for any type of biologic measurement that involves a data transformation step commonly used to reduce nuisance variation before analysis.

References

Bland, J. M. and D. G. Altman (1986). Statistical method for assessing agreement between two methods of clinical measurement. *The Lancet*, 307–310.

Bolstad, B., R. A. Irizarry, M. Astrand, and T. P. Spend (2003). Exploration, normalization,

and summaries of high density oligonucleotide array probe level data. *Bioinformatics 19*, 185–193.

Box, G. E. and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, 244–243.

Cui, X., M. K. Kerr, and G. A. Churchill (2003). Transformations for cdna microarray data. *Statistical Applications in Genetics and Molecular Biology 2*, 1–20.

Gentleman, R., T. Rossini, S. Dudoit, and K. Hornik (2003). The bioconductor faq. *[http://www.bioconductor.org/]*.

Huber, W., A. von Heydebreck, H. Morgan, A. Poustka, and M. Vingro (2003). Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology 2 2*.

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

Kadane, J. B. (1978). A comment on "normalization in point estimation" (stma V20 303). *Journal of Econometrics 7*, 123–125.

Li, C. and W. H. Wong (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science 98*, 31–36.

Mardia, K. V., J. T.Kent, and J. M. Bibby (1979). *Multivariate Analysis*. Academic.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics Supplement 32*, 496–501.

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

Schadt, E. E., C. Li, B. Ellis, and W. H. Wong (2000). Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry 80*, 192–202.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

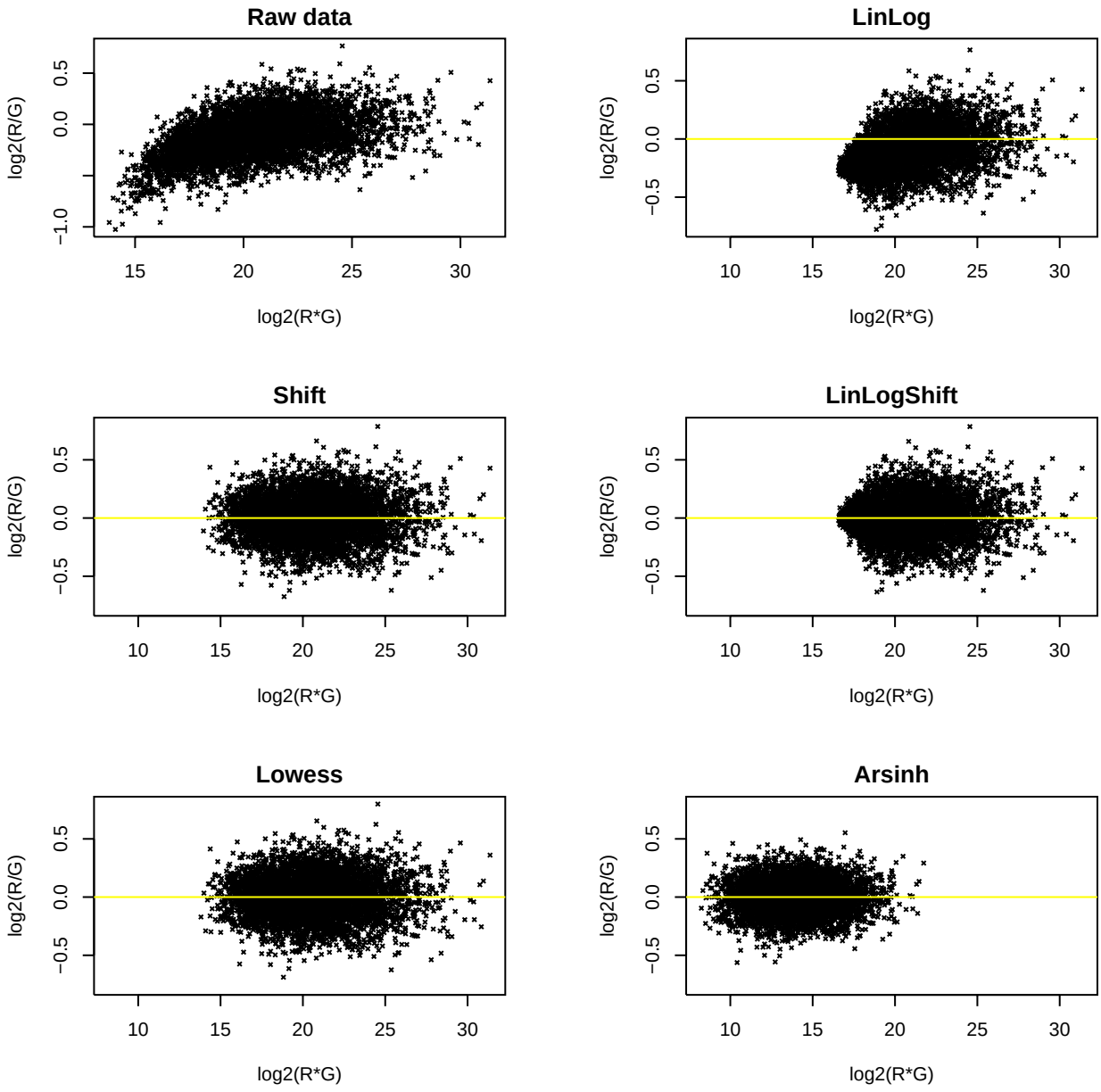

Figure 1: One Array with Background Difference

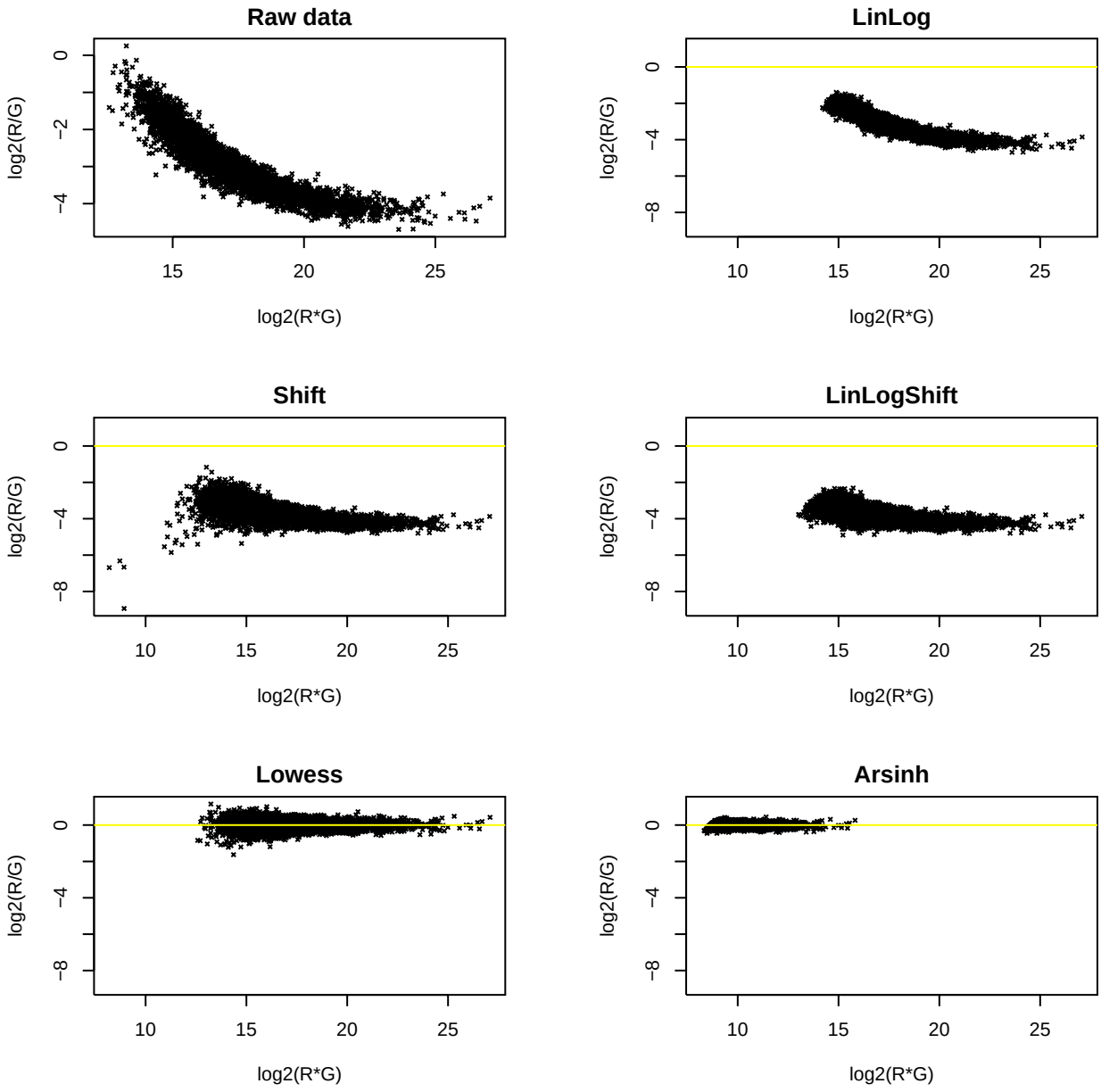

Figure 2: One Array with Slope Difference

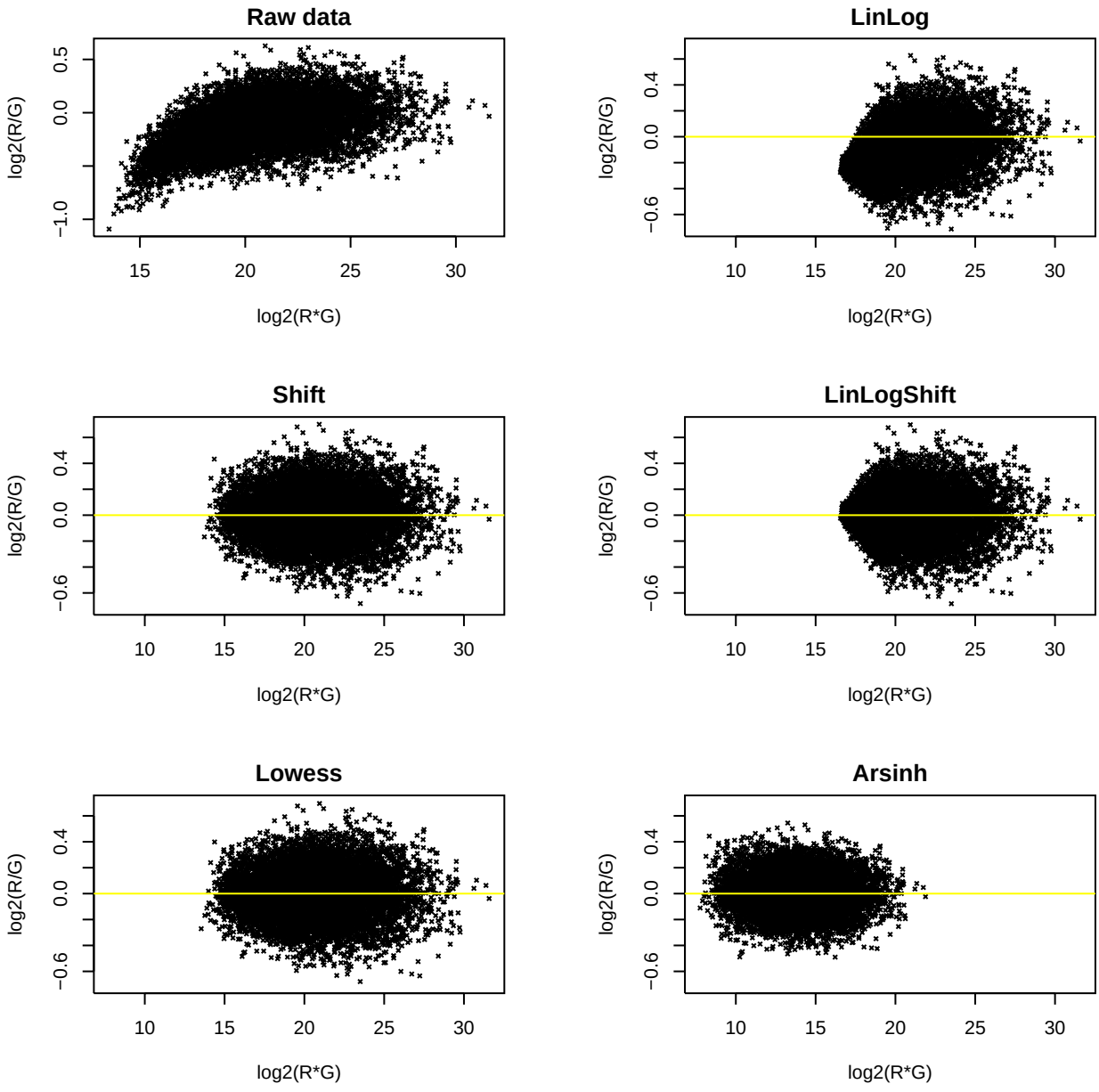

Figure 3: Two Arrays with Background Difference

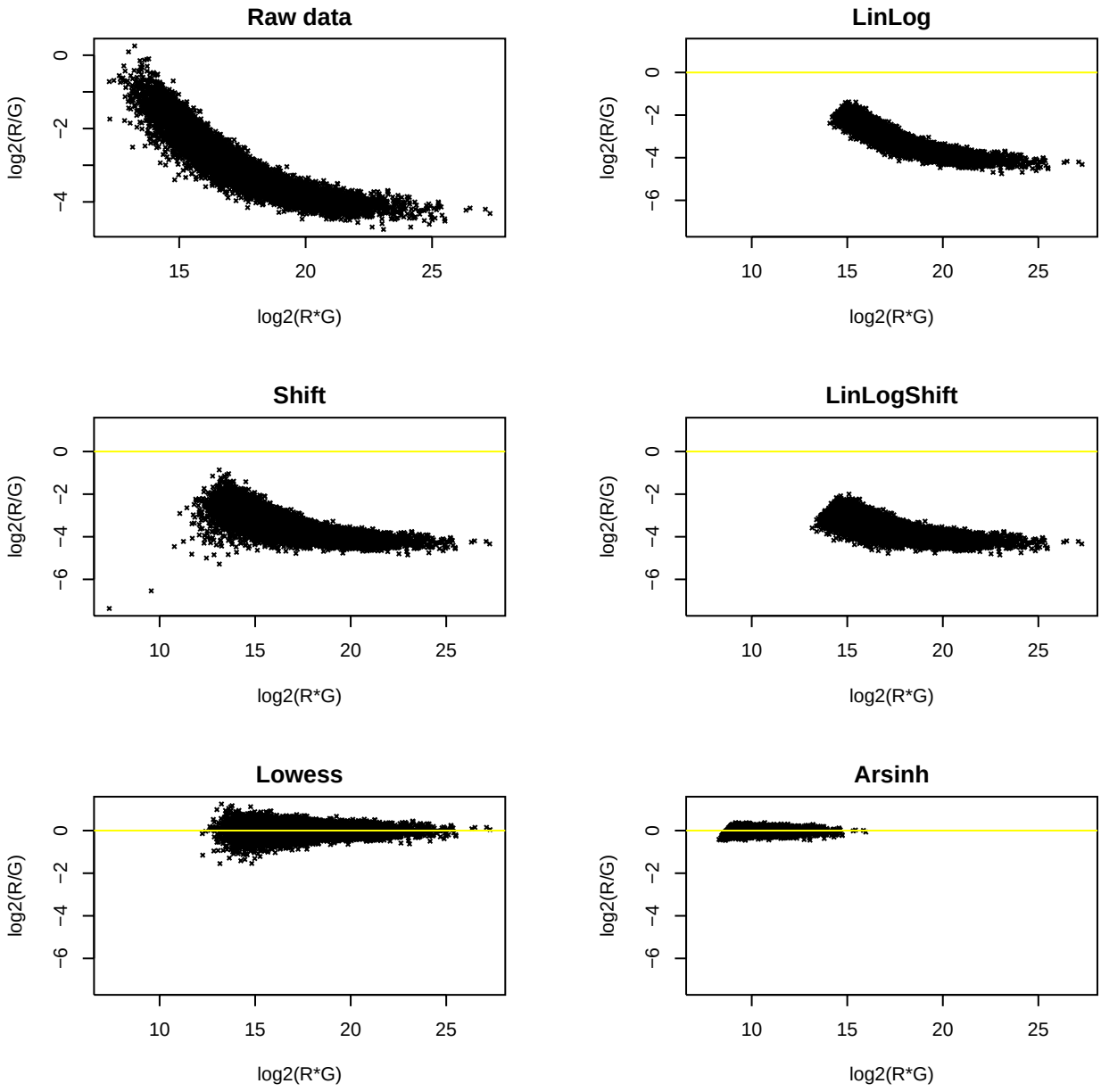

Figure 4: Two Arrays with Slope Difference