

# Transformation and Normalization of Oligonucleotide Microarray Data

Sue C. Geller  
Department of Mathematics  
Texas A&M University  
College Station, TX 77843-3368\*

Jeff P. Gregg  
Department of Pathology, School of Medicine  
University of California, Davis,  
Sacramento, CA 95817

Paul Hagerman  
Department of Biological Chemistry, School of Medicine  
University of California, Davis,  
Davis, CA 95616

David M. Rocke  
Department of Applied Science and Division of Biostatistics  
University of California, Davis,  
Davis, CA 95616

11 July, 2002

---

\*To whom correspondence should be addressed

## Abstract

**Motivation:** Most methods of analyzing microarray data or doing power calculations have an underlying assumption of constant variance across all levels of gene expression. The most common transformation, the logarithm, results in data that have constant variance at high levels but not at low levels. Rocke and Durbin (2001) showed that data from spotted arrays fit a two component model and Durbin, Hardin, Hawkins, and Rocke (2002) and Huber et al. (2002) provided a transformation that stabilizes the variance as well as symmetrizes and normalizes the error structure. We wish to evaluate the applicability of this transformation to the error structure of GeneChip microarrays.

**Results:** We show that data from Affymetrix GeneChips conform to the same two component model as in Durbin, Hardin, Hawkins, and Rocke (2002) and Huber et al. (2002), so that their transformation is appropriate. In addition we provide a method for normalization of spotted arrays and Affymetrix GeneChips simultaneous with the determination of the transformation, producing a data set without chip or slide effects but with constant variance and with symmetric errors. This transformation/normalization process can be thought of as a machine calibration in that it requires a few biologically-constant replicates of one sample to determine the constant needed to specify the transformation and normalize. It is hypothesized that this constant needs to be found only once for a given technology in a lab, perhaps with periodic updates. It does not require extensive replication in each study. Furthermore, the variance of the transformed pilot data can be used to do power calculations using standard power analysis programs.

**Contact:** geller@math.tamu.edu

**Availability:** SPLUS code for the transformation/normalization for four replicates is available from the first author upon request.

**Keywords:** Transformation, normalization, power calculations, microarray, Affymetrix GeneChip, genomics, statistical analysis, autism.

# 1 Introduction

Appropriate ways of preprocessing microarray data before analysis is a topic of continuing interest and discussion. Among the reasons for such preprocessing are normalization, i.e., removing slide/chip effects, background intensities, and other sources of systematic error, and transforming the data so that assumptions needed for the analysis are met. The order in which normalization and transformation are performed varies as well, but the two topics are usually treated separately.

It was shown in Rocke and Durbin (2001) that the data from spotted microarrays conformed to a two-component model of error, i.e.,  $y = \alpha + \mu e^\eta + \epsilon$ , where  $y$  is the measured intensity,  $\mu$  is the (unknown) expression level in arbitrary units, and  $\eta$  and  $\epsilon$  are normally-distributed error terms with mean 0 and variance  $\sigma_\eta^2$  and  $\sigma_\epsilon^2$ , respectively. Note that the additive error dominates when  $\mu$  is small, and the proportional error dominates when  $\mu$  is large. They described a procedure by which to estimate  $\sigma_\eta$  and  $\sigma_\epsilon$ , the former being approximately the standard deviation of the high level log-transformed data and whose estimation requires biologically-constant replication of at least one sample. Durbin, Hardin, Hawkins, and Rocke (2002) and Huber et al. (2002) independently developed a data transformation to stabilize the variance for data that fit the two-component model, namely,  $z = \ln\left((y - \alpha) + \sqrt{(y - \alpha)^2 + c}\right)$ , where  $c$  is a constant to be determined from the data, specifically  $c = \hat{\sigma}_\epsilon^2 / \hat{\sigma}_\eta^2$ . They noted that the commonly used log transform stabilizes the variance for highly expressed genes but not for genes with low or zero expression, and showed that the transformation was effective for spotted microarray data. The transformed data not only has constant variance but the errors and transformed data are approximately symmetric.

Many methods of normalization, such as subtraction of an estimated background intensity and the commonly used average difference score for Affymetrix GeneChips can result in negative measured intensities. This can also happen with other normalization methods such as simple ANOVA or regression techniques, or the program dChip, which uses an iterative procedure with regression and outlier detection (<http://www.dchip.org>; Li and Wong 2001). The negative values for intensities (e.g., when MM > PM) cause problems when the commonly used log transform is applied to the normalized data. Also, if the variance is not constant across expression levels, then the

mean chip/slide intensity – the overall mean intensity for all the chips/slides is not an optimal measure of the chip/slide effect.

We will demonstrate that data from Affymetrix GeneChips conform at least approximately to the two-component model of error. Then we shall provide a procedure by which to find the transformation constant  $c$  and normalize the data at the same time. This simultaneous procedure is necessary under the assumptions of our model. This method avoids many of the problems of alternative normalization and transformation methods. In particular there is no problem with negative observed intensities.

## 2 Data

In order to investigate replicability and the appropriateness of the two-component model of error for data from Affymetrix chips, a single lymphoblastoid cell line from one autistic child was grown in four separate T75 flasks. The cells were split from the same parent flask and grown in the same incubator, same shelf, at the same time. Four separate RNA extractions were performed when the cells were near confluence. cDNA synthesis and in vitro transcription (IVT) labeling were performed on each sample. Each sample was then hybridized to an Affymetrix Human Genome U95Av2 oligonucleotide GeneChip array, which contains 12,625 probe sets, with each probe set designed to represent a single human gene. Note that these are true replicates of the whole measurement process, not just machine replicates in which the same sample after processing is divided and hybridized to several different chips.<sup>1</sup>

This study is part of a larger project to investigate the genetic basis of autism, a behavioral diagnosis that reflects complex and ill-defined patterns of inheritance and possible environmental factors. Although autism is highly heritable, with over 90% concordance for autism spectrum disorders, genome-wide linkage studies have been unable to define specific chromosomal loci for the preponderance of autistic individuals, excluding known genetic disorders

---

<sup>1</sup>There are a number of differing types of replicates depending on where in the process one starts. A fundamental assumption of our model is that the replicates used to estimate it come from the same biological sample for all replicates in a group. Thus, we are considering measurement error, not variability within and between organisms. The samples we are using for this paper all come from the same cell line. We are measuring, in this case, all of the sources of error beginning with the collection of the cells, through hybridizing to the chip and reading the results.

(e.g., Rett syndrome, fragile X syndrome, tuberous sclerosis, 15q(dup)) that have autistic features as a variable component of their phenotypes. An alternative approach to the identification of genetic components of autism, including those conferring differential susceptibility to putative environmental factors, is the identification of patterns of altered gene expression through microarray methods. This approach entails the analysis of blood and tissue samples from autistic patients and unaffected family members with the objective of identifying patterns of gene expression that associate with the autism spectrum phenotype. Information derived from such investigations would be of critical importance for both predictive efforts (which children are at increased risk to develop autism) and efforts to define the various biochemical/genetic mechanisms that lead to autism.

All graphs and computed statistics in this paper are from the data set of Perfect Match (PM)/ Mismatch (MM) differences for 12625 genes for 4 replicates or flasks. We used the average difference (PM – MM) data as they are reported by Microarray Suite 4.0 (MAS 4.0; Affymetrix Inc.) without using their normalization program. Version 4 reports negative values when the MM is greater than the PM.<sup>2</sup> Since the standard deviation of the low expression level data is approximately 8, we considered genes for which the median value was less than or equal to  $-25$  to be unusable outliers,<sup>3</sup> and eliminated those 655 genes. In general the data can be thought of as an  $n \times m$  array of intensity levels where  $n$  is the number of genes,  $m$  the number of slides or chips analyzed, and  $n \gg m$ ,  $11970 \times 4$  in our example data.

### 3 Two-Component Model

Figure 1 shows that the data fit the two-component model. It is a graph of the median of the four replicates for each of the 11970 genes by the interquartile range for that gene. The solid line is the loess smooth of the scatter plot, whose slope for large positive intensities shows that the scale of the error is linear in the location, as would be expected. Due to the large range of

---

<sup>2</sup>Version 5 avoids negative average difference scores by artificially constraining the results to be positive. It is not yet clear whether this makes analysis more or less difficult.

<sup>3</sup>Under the usual assumptions about PM – MM, the expected value should never be less than zero. Thus, very few data points should be less than three standard deviations below zero, and even fewer medians of four should be less than this cutoff. By eliminating data in which the median PM – MM is less than  $-25$ , we are eliminating data that do not conform the expected behavior for expressed or unexpressed genes.

medians, the constant variance for intensities near 0 is not visible. Note that we use median and interquartile range (IQR) instead of mean and standard deviation for robustness against outliers, which are not uncommon in this type of data. With four observations, we use the difference of the middle two numbers as the IQR. The Appendix contains definitions of the finite-sample IQR for sample sizes up to 20, and the required constants for estimating the standard deviation from the IQR.

Figure 2 is a similar scatter plot for the natural log transformed data. Note that the variance is approximately constant only for large expression levels and that there are many outliers or high variance data at lower levels. Furthermore, only 8339 of the 12625 genes (or 11970 genes of the data subset) were usable as the other approximately 30% of the data had one or more of the replicates with non-positive values. Omission of such a large amount of data is not optimal. Even if it is considered that these data are unimportant in this sample, since the genes are not expressed at a high level, comparison of different conditions is made much more difficult if a gene is expressed at a low level in one sample and a high level in another. Simply replacing negative estimated expression values by some arbitrary positive value such as 10 or 20, or by an estimated mean, is also non-optimal since it distorts the variability pattern of the data.

## 4 Transformation and Normalization

The approach we pursue is to formulate a model that fits the variance patterns in the data, and that contains normalization constants, and then use a procedure that can simultaneously determine the transformation parameter and the normalization. The statistical model used is that, for gene  $i$  and chip  $j$ , the measured (average difference) value  $y_{ij}$  satisfies

$$f_c(y_{ij}) = \mu_i + n_j + \epsilon_{ij}$$

where  $\mu_i$  is the true expression of gene  $i$  in the sample,  $n_j$  is an additive chip normalization, and  $\epsilon_{ij}$  is an additive symmetric measurement error.

The back-transformed, normalized measurements

$$\tilde{y}_{ij} = f_c^{-1}(f_c(y_{ij}) - n_j + \bar{n}) + \alpha$$

are assumed to fit the two-component model, so that

$$V(\tilde{y}_{ij}) = \sigma_\epsilon^2 + S_\eta^2 \mu_i^2.$$

In this model,  $S_\eta^2 = e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1)$ , which is to the first order the same as  $\sigma_\eta^2$ . For example, if  $\sigma_\eta = .1$ , corresponding to a 10% coefficient of variation, then  $S_\eta = .1008$ .

The transformation used is

$$f_c(y) = \ln \left( y - \alpha + \sqrt{(y - \alpha)^2 + c} \right)$$

where  $c = \sigma_\epsilon^2/S_\eta^2$ , or equivalently to the first order  $c = \sigma_\epsilon^2/\sigma_\eta^2$ . Note that we are treating  $\alpha$  as negligible given that the primary data are PM-MM, which can be presumed to have a mean near zero for unexpressed genes. We can estimate  $\sigma_\epsilon^2$  from the low level data and  $\sigma_\eta^2$  from the high level data, as in Rocke and Durbin (2001), but the values of  $\hat{\sigma}_\eta$ ,  $\hat{\sigma}_\epsilon$ , and  $c$  are interdependent in the sense that we cannot determine  $\hat{\sigma}_\eta$  and  $\hat{\sigma}_\epsilon$  without knowing  $c$ , and vice versa. We solve this problem using the following iteration:

1. Starting with a trial value of  $c$ , transform the data using  $f_c(y_{ij}) = \ln \left( y_{ij} - \hat{\alpha} + \sqrt{(y_{ij} - \hat{\alpha})^2 + c} \right)$ .
2. Determine the normalization constants  $n_j$  taking the median of all genes over each slide and subtracting those from the individual transformed gene expression values, then adding the median of all the genes on all the slides to keep the average level the same.
3. Back transform the normalized transformed values using the inverse transformation

$$f_c^{-1}(z) = \frac{e^z - ce^{-z}}{2} + \hat{\alpha}$$

4. Determine the parameters  $\sigma_\epsilon$  and  $\sigma_\eta$  of the two-component model by pooling low and high levels of expression as given in the Appendix. Ordinarily, we would also re-estimate  $\alpha$ , but in this case, we are assuming  $\alpha = 0$ .
5. Determine a new transformation parameter  $c = \sigma_\epsilon^2/\sigma_\eta^2$ , and return to step 1 (using the original data not the back-transformed data).
6. Stop the process when the parameter values stop changing.

For this application, we have assumed that the average difference scores of unexpressed genes average out to near zero, so that  $\alpha$  is assumed to be

zero, rather than estimated. If this could not be assumed, the intensities would have to be background subtracted at each step before transformation (Durbin et al. 2002). Also note that we cannot determine the transformation accurately without the iterative normalization. In our data the flask effects (flask median minus overall median) on the transformed scale ranged from  $1/4$  to  $3/4$  of a robust estimate of the standard deviation of the errors, so would have affected our computation of  $c$ .

The above procedure was applied to the autism data. It converged after six iterations to  $c = 4866$  (from  $\sigma_\epsilon = 8.752$  and  $\sigma_\eta = .1255$ ). Figure 3 is a scatter plot of the median and IQR of the transformed data ( $c = 4866$ ) from which the flask effects had first been subtracted. Note that the loess line is very close to constant. The errors are symmetric but long-tailed compared to a normal distribution.

The procedure is not sensitive to the initial value of  $c$ . For example, the iteration starting with  $c = 50,000$  for the above data set converged in ten iterations to the same solution, compared to six or eight iterations using better starting points. When we used the full 12625 genes, we found the value of  $c$  to be more sensitive to the method of pooling to calculate the variance parameters, in particular to the choice of how many IQRs or standard deviations to pool. Using too few replicate genes gives a more variable estimate, while using too many may result in bias. We pooled 125 for each of  $\sigma_\eta$  and  $\sigma_\epsilon$ . For the full data set, pooling 125 for each resulted in  $c_{125} = 4617$ , whereas pooling 200 for each resulted in  $c_{200} = 5013$ , a change of under 10%. By eliminating the genes with very negative median values, a choice made because such negative values are suspect as to their being due simply to random error,  $c$  was not as sensitive to the number of IQRs pooled, since  $c_{125} = 4866$  and  $c_{200} = 4842$ , a change of under 1%. Since high intensity levels fall off more sharply than low ones increase, in general a choice of pooling between 100 and 150 is recommended for estimating  $\sigma_\eta$  and between 100 and 200 for estimating  $\sigma_\epsilon$  (for this particular chip), but looking at the particular data set and being sure that one is pooling very large and very close to zero is a better way to pick the number to pool for  $\sigma_\eta$  and  $\sigma_\epsilon$ , resp.

## 5 Power Calculations

Since the transformed data are approximately symmetric, with approximately constant variance across intensity levels, one can use the variance of the transformed data and information on variability between biological specimens (e.g., from pilot data) to do power calculations using standard software such as that at [www.swogstat.org/stat/public/default.htm](http://www.swogstat.org/stat/public/default.htm) or using packages such as nQuery.

## 6 Conclusion

The iterative technique produces a transformed data set with the following nice properties:

- constant variance of replicates for different genes,
- symmetric errors, and
- no systematic differences from slide to slide.

The normalization and transformation process adds 1 to 3 chips to an experiment, since biologically identical cells must be used as replicates for the calibration. However, our process removes the requirement of duplicating every sample, a time and money saver.

The transformed data are appropriate to use in analysis employing both non-parametric and parametric methods. Furthermore, standard software can then be used for power calculations.

## Acknowledgments

The research reported in this paper was supported by grants from the National Science Foundation (ACI 96-19020, and DMS 98-70172), the National Institute of Environmental Health Sciences, National Institutes of Health (5-P42 ES04699-16 with funding provided by EPA and 1-P01-ES11269), and by the UC Davis MIND Institute Biomarkers Initiative. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIEHS, NIH, EPA, or NSF.

## References

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

Li, C. and Wong, W. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci.*, 98, 31-36.

Rocke, D.M. (1992)  $\bar{X}_Q$  and  $R_Q$  Charts: Robust Control Charts, *The Statistician*, 41, 97-104.

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

## Appendix: Fitting the Two-Component Model

In this paper, we use a variant of the procedure of Rocke and Durbin (2001), the main differences being the use of the median and IQR instead of the mean and standard deviation. In addition, we use a fixed number of genes to pool at the high and low expression ends. The following is description of the procedure to estimate  $\hat{c}$ ,  $\sigma_\epsilon$ , and  $\sigma_\eta$  after an initial estimate of  $c$  has been used to transform the data and remove flask effects.

Let  $\tilde{y}_{ij}$  be the normalized, back-transformed value for gene  $i$  ( $1 \leq i \leq n$ ) on chip  $j$  ( $1 \leq j \leq m$ ) as input for step 4 of the algorithm given in Section 4. Let  $M_i$  be the median over the  $m$  replicates of the values for gene  $i$ ; let  $\tilde{M}_i$  be the median of the natural logarithms of the replicates (which may differ from  $\ln(M_i)$  if  $m$  is even) when all replicates are positive; and let  $s_i$  be an estimate of the scale of the replicates for gene  $i$  based on the IQR and defined as follows:

- Let  $z_{(1)}, z_{(2)}, \dots, z_{(m)}$  be the replicates for a gene sorted from smallest to largest.

Table 1: Constants for determination of estimates of scale based on the IQR.

m	a	b	$d_2^Q$
2	1	2	1.1284
3	1	3	1.6926
4	2	3	0.5940
5	2	4	0.9900
6	2	5	1.2835
7	2	6	1.5147
8	3	6	0.9456
9	3	7	1.1439
10	3	8	1.3121
11	3	9	1.4577
12	4	9	1.0737
13	4	10	1.2057
14	4	11	1.3235
15	4	12	1.4298
16	5	12	1.1400
17	5	13	1.2389
18	5	14	1.3296
19	5	15	1.4132
20	6	15	1.1806

- Let  $s_i = (z_{(b)} - z_{(a)})/d_2^Q$ , where  $a$ ,  $b$ , and  $d_2^Q$  are given in Table 1. If  $m > 20$ , use  $a = \lceil m/4 \rceil$ ,  $b = m - a + 1$ , and  $d_2^Q = 1.35$ . Alternatively, more exact values for  $d_2^Q$  for  $m > 20$  can be derived from tables of normal order statistics as described in Rocke (1992)

Similarly, let  $\tilde{s}_i$  be an estimate of the scale of the logarithms of the replicates defined as for  $s_i$ . If there are differing numbers of replicates due to missing values or other considerations, one uses the actual number of replicates for each group. This allows the method to cope with missing values.

Let  $h \ll n/2$  and  $\ell \ll n/2$ ; let  $H$  be the set of indices  $i$  such that  $\{\tilde{M}_i | i \in H\}$  are the  $h$  largest medians of the logarithms of the data, and let  $L$  be the set of indices  $i$  such that  $\{M_i | i \in L\}$  are the  $\ell$  smallest medians of

the raw data. Our estimates of the two-component model parameters are

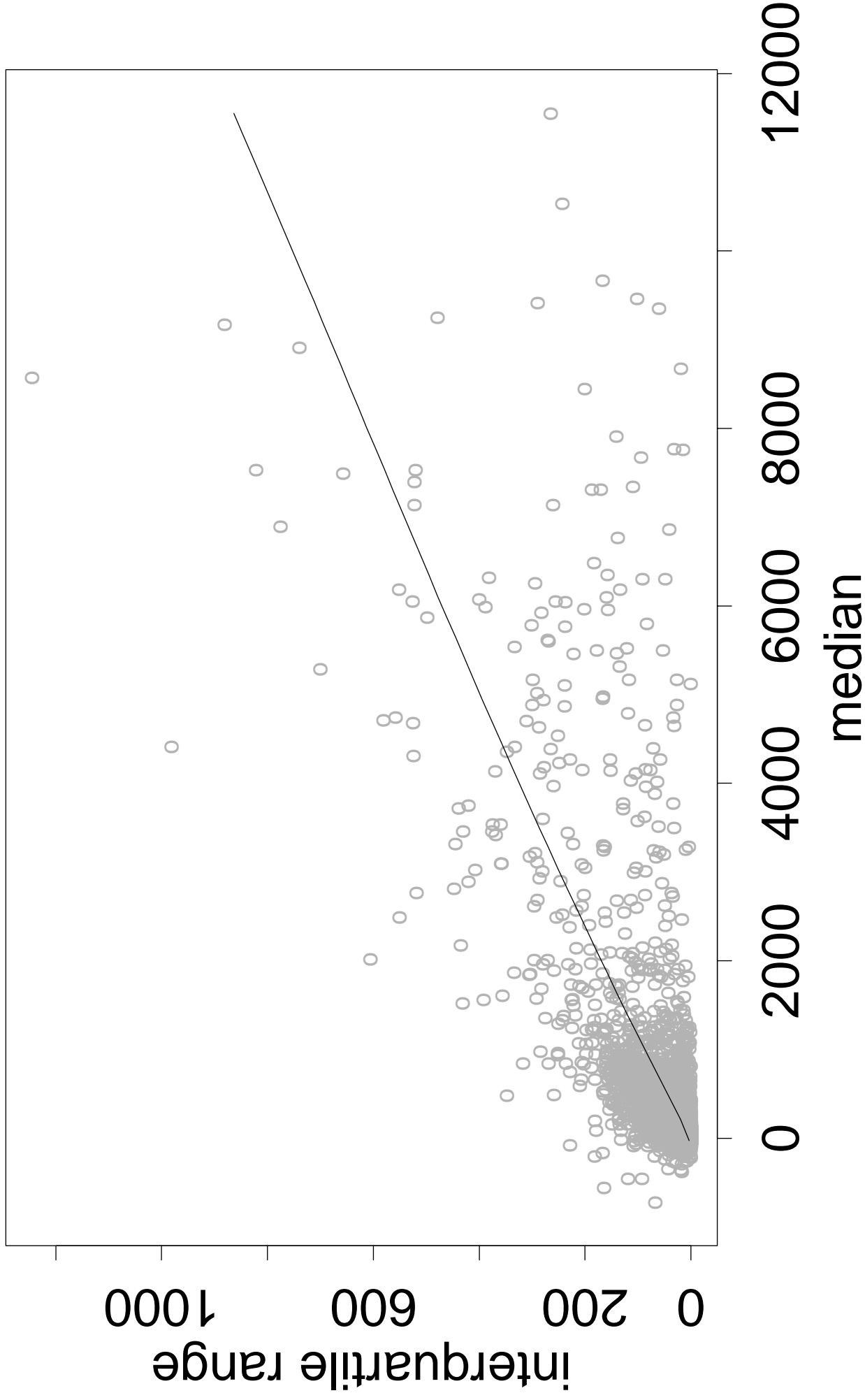
$$\hat{\alpha} = \ell^{-1} \sum_{i \in L} M_i \quad (1)$$

$$\hat{\sigma}_\epsilon = \ell^{-1} \sum_{i \in L} s_i \quad (2)$$

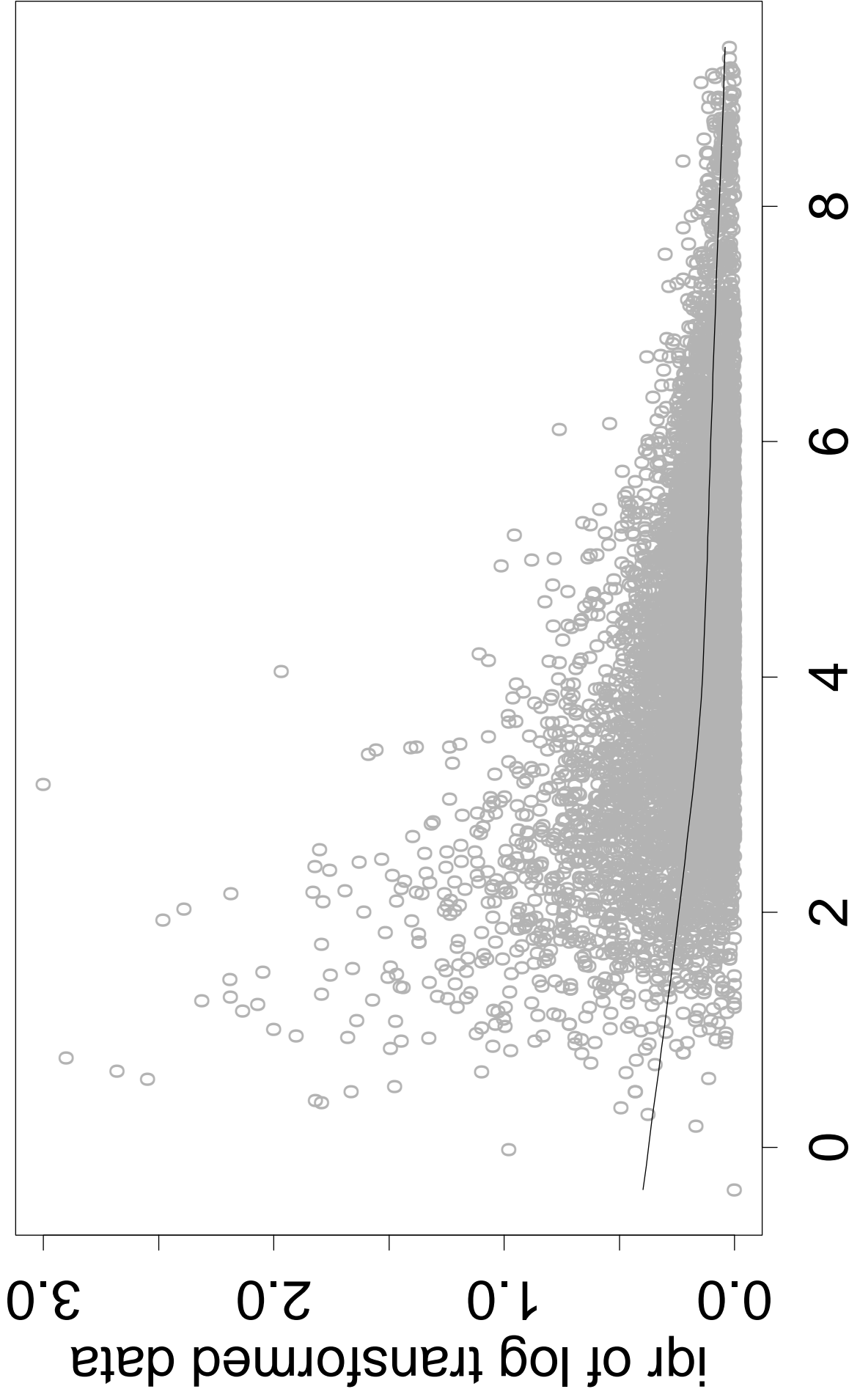
$$\hat{\sigma}_\eta = h^{-1} \sum_{i \in H} \tilde{s}_i \quad (3)$$

In the computations reported in this paper, we have assumed that  $\alpha = 0$ , so have not estimated  $\alpha$  at each step.

# Untransformed Data



# Log Transformed Data



median of log transformed data  
8339 of 12625 data points usable

# Transformed Data with no Flask Effects

