Measurement Error and Data Transformation for Gene Expression, Proteomics, and Metabolomics Data

David M. Rocke

Department of Applied Science, College of Engineering,

Division of Biostatistics, School of Medicine, and

Center for Image Processing and Integrated Computing

University of California, Davis

IMA

September, 2003

Overview

- Gene expression data from microarrays present many challenging problems for analysts. The data exhibit complicated error structures which are not widely recognized. The dimension of the data is usually much higher than the sample size.

- We present a model for measurement error in gene expression data that explains a number of problems currently facing users of these data.

- We present a class of data transformations specifically tuned to microarray data (and other high-throughput assay data) that can stabilize the variance and allow more effective use of standard statistical methods. We call these the generalized logarithmic (glog) transformations after Munson.

- We show how the transformation parameter(s) can be estimated either from a few replicate arrays or using maximum likelihood in the context of a linear model analysis.

- We show how the remaining variance heterogeneity can be accounted for with a hierarchical Bayes model that can be easily estimated in an empirical Bayes fashion.

Gene Expression Arrays

- Genes produce proteins with an intermediary of mRNA. At any given time, most genes in a cell are not expressed (translated to proteins). Some genes are expressed only during development, others are specific to tissue types or environmental condition.

- There is also variation in genetic structure between organisms or between tissues in the same organism in case of mutation.

- The purpose of gene expression arrays is to measure the gene expression in the given tissue/cell at the given time by measuring concentration of mRNA for many genes at the same time.

- In one version of the technology, microscopic spots containing cDNA clones or synthesized oligonucleotides are deposited on a glass slide using a micropipette from a supply in wells of microplates.

- The sample is reverse-transcribed, labeled with a fluorescent dye, hybridized to the spots on the slide, and the intensity of fluorescence measured with a laser scanner.

- In two-color arrays, treatment and control samples (or any two samples) are labeled with two color dyes, mixed, and hybridized together.

- There are other variations including one-color cDNA arrays and the oligonucleotide arrays such as those from Affymetrix, but many of the statistical issues are the same across technologies.

Measurement Error

Some well known properties of measurement error in gene expression microarrays (GEM's) include the following:

- For high gene expression, the standard deviation of the response is approximately proportional to the mean response, so that the CV is approximately constant.

- For low levels of expression, the CV is much higher.

- Expression is commonly analyzed on the log scale, so that for high levels the SD is approximately constant, but for low levels of expression it rises.

- Comparisons of expression are usually expressed as n-fold, corresponding to the ratio of responses, of which the logarithm would be well behaved, but only if both genes are highly expressed.

- These phenomena occur in many measurement technologies, but are more important in high-throughput assays like GEM's.

- What is the fold increase when a gene goes from zero expression in the control case to positive expression in the treatment case?

- Which is biologically more important: an increase in expression from 0 to 100 or an increase from 100 to 200?

- Because the genes expressed on a GEM are a mixture of levels of expression, and a large number have no measurable expression, the average standard deviation of the logarithms across all genes is quite high. In reality, the standard deviation of the logs is often quite low for genes expressed well above background.

Error Model for Gene Expression Microarrays

The error model we use that motivates the data transformation is as follows:

$$y = \alpha + \mu e^{\eta} + \epsilon$$

where y is the intensity measurement, μ is the expression level in arbitrary units, and α is the mean background (mean intensity of unexpressed genes). Our best estimate of μ is $y - \hat{\alpha}$, the background-corrected intensity.

Under this model, the variance of the background-corrected response $y - \alpha$ at concentration μ is given by

$$\text{Var}(y - \alpha) = \mu^2 S_\eta^2 + \sigma_\epsilon^2.$$

where

$$S_\eta = \sqrt{e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1)},$$

which is of the form

$$\begin{aligned} E(z) &= \mu \\ V(z) &= a^2 + b^2\mu^2 \end{aligned}$$

It can be shown that

$$\text{Var}\{\ln(y-\alpha)\} \approx \sigma_\eta^2 + \sigma_\epsilon^2/\mu^2.$$

and

$$\text{Var}\{\ln(y)\} \approx \frac{\mu^2\sigma_\eta^2 + \sigma_\epsilon^2}{(\mu+\alpha)^2} \approx \frac{\mu^2\sigma_\eta^2 + \sigma_\epsilon^2}{y^2}.$$

Note the implication for use of logarithms on background-corrected data.

Data Transformation

- Logarithms stabilize the variance for high levels, but increase the variance for low levels.

- Log expression ratios have constant variance only if both genes are expressed well above background.

Transformation and Variance

Let y_i estimate μ_i, and suppose that

$$\mathrm{Var}(y_i) = \sigma_0^2 v(\mu_i).$$

Consider a transformation $z = f(y)$. It is well known that, up to the first order,

$$\mathrm{Var}(z_i) = (f'(\mu_i))^2 \sigma_0^2 v(\mu_i).$$

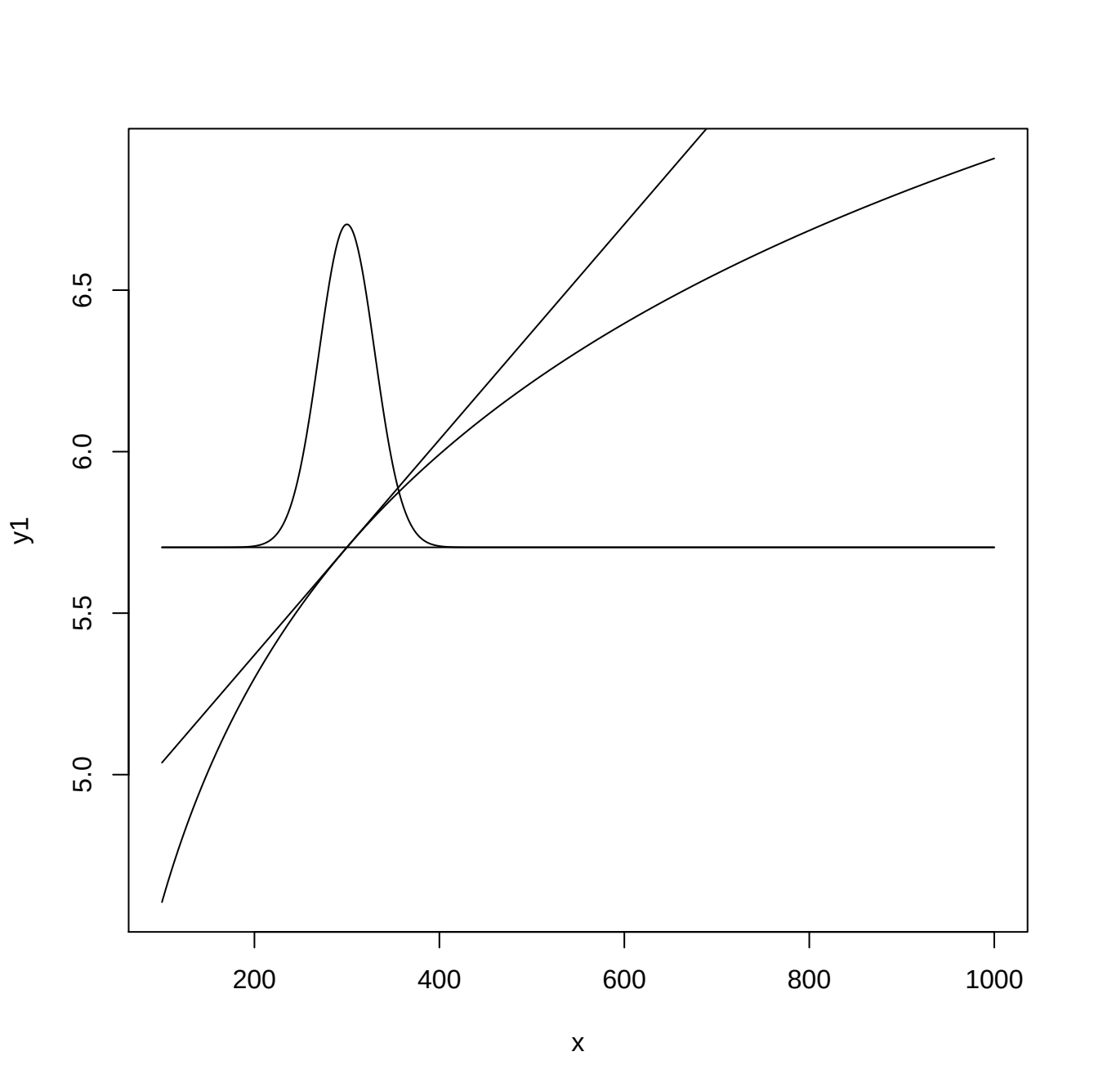
This is called *propogation of error* or the *delta method*.

This is based on a linear approximation to the transformation function. If

$$\mathrm{SD}(y) = \sigma,$$

and if $z = a + by$, then

$$\mathrm{SD}(z) = b\sigma,$$


y1
6.5
6.0
5.5
5.0
200
400
600
800
1000
x

Thus a transformation that fully stabilizes the variance would be one in which

$$(f'(\mu_i))^2 = \frac{1}{\sigma_0^2 v(\mu_i)} \quad \text{or}$$

$$f'(\mu_i) = \frac{1}{\sqrt{\sigma_0^2 v(\mu_i)}}$$

If $v(\mu) = \mu^2$, then

$$f'(\mu) = \frac{1}{\mu}.$$

so

$$f(\mu) = \ln(\mu).$$

(as is well known).

If $v(\mu) = \mu$ (as in Poisson data), then

$$f'(\mu) = \mu^{-1/2}.$$

so

$$f(\mu) = \sqrt{\mu}.$$

(as is also well known).

In much of chemical analysis and biological measurement data, a reasonable model is

$$y = \mu e^{\eta} + \epsilon,$$

so that

$$V(y) = a^2 + b^2\mu^2$$

where $a = \sigma_\epsilon$ and $b = S\eta$.

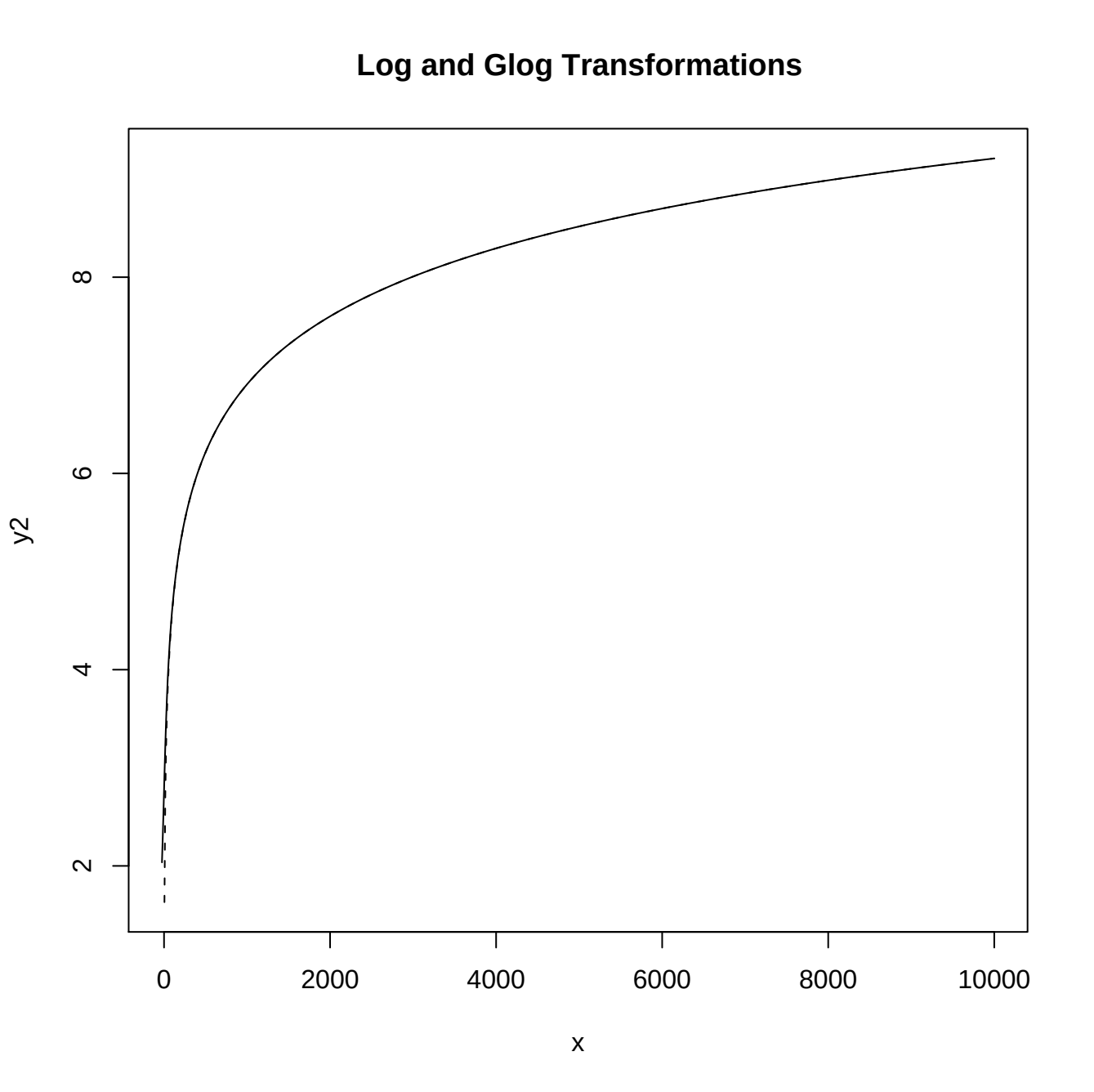
With this variance function, we have

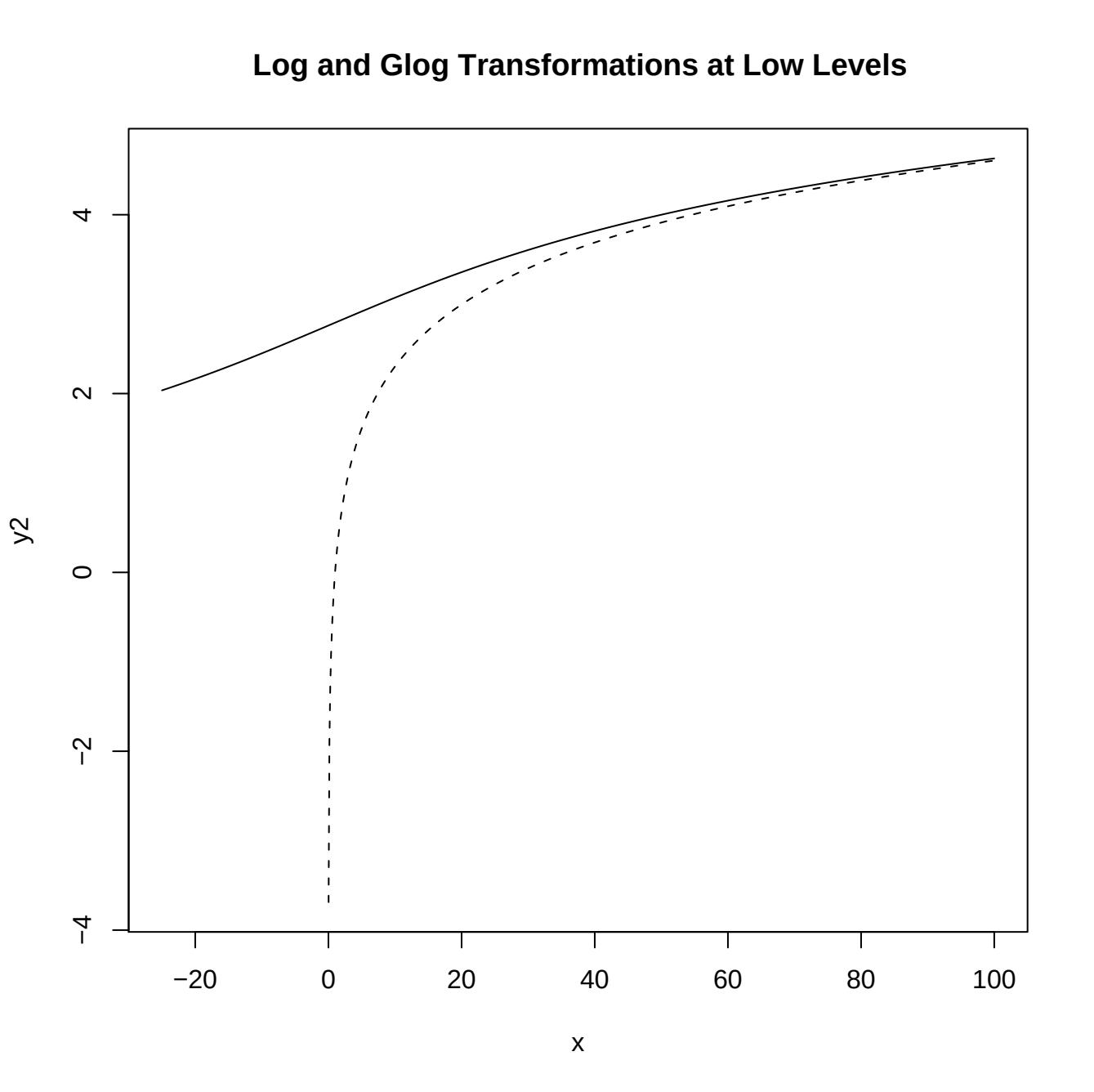
$$f'(\mu) = \frac{1}{\sqrt{a^2 + b^2\mu^2}}.$$

which integrates to what we call the generalized log (glog) function

$$f(\mu) = \ln(\mu + \sqrt{\mu^2 + a^2/b^2}).$$

(Durbin, Hardin, Hawkins, and Rocke 2002; Hawkins 2002; Huber, von Heydebreck, Sültmann, Poustka, and Vingron 2002; Munson 2001)


Log and Glog Transformations
y2
2
4
6
8
0
2000
4000
6000
8000
10000
x


Log and Glog Transformations at Low Levels
y2
4
2
0
−2
−4
−20
0
20
40
60
80
100
x

To use this, we must estimate a, the standard deviation of the untransformed data at low levels, and b, which is the standard deviation of the logged data at high levels. Alternatively, we can estimate directly the transformation parameter $\lambda = a^2/b^2$. We also need to estimate the parameter α in the TCM, either separately or together with λ. The glog transform is then

$$h_{\lambda,\alpha}(y) = \ln\left(y - \alpha + \sqrt{(y-\alpha)^2 + \lambda}\right).$$

This helps solve the puzzle of comparing a change from 0 to 50 to a change from 200 to 300. Suppose that the standard deviation at 0 is 10, and the high-level CV is 10%. Then

- A change from 0 to 50 is five standard deviations ($5 \times 10 = 50 = 50 - 0$).

- A change from 200 to 300 is also five standard deviations ($300/200 = 150\% = 5 \times 10\%$).

- So is a change from 1000 to 1500 ($1500/1000 = 150\% = 5 \times 10\%$).

- The biological significance of any of these is unknown. Different transcripts can be active at vastly different levels.

- But the glog transformation makes an equal change equally statistically significant.

Determining Differentially Expressed Genes

Consider an experiment on four types of cell lines A, B, C, and D, with two samples per type, each of the eight measured with an Affymetrix U95A human gene array. Let y_{ijk} be the measured expression for gene i in group j and array k in group j. We used MAS 4.0 Average Difference since it does not artificially compress low-level data.

Steps in the Analysis

- Background correct each array so that 0 expression corresponds to 0 signal.

- Transform the data to constant variance using a suitably chosen glog or alternative transformation (started log, hybrid log).

- Normalize the chips additively.

- The transformation should remove systematic dependence of the gene-specific variance on the mean expresssion, but the gene-specific variance may still differ from a global average. Estimate the gene-specific variance using all the information available.

- Test each gene for differential expression against the estimate of the gene-specific variance. Obtain a p-value for each gene.

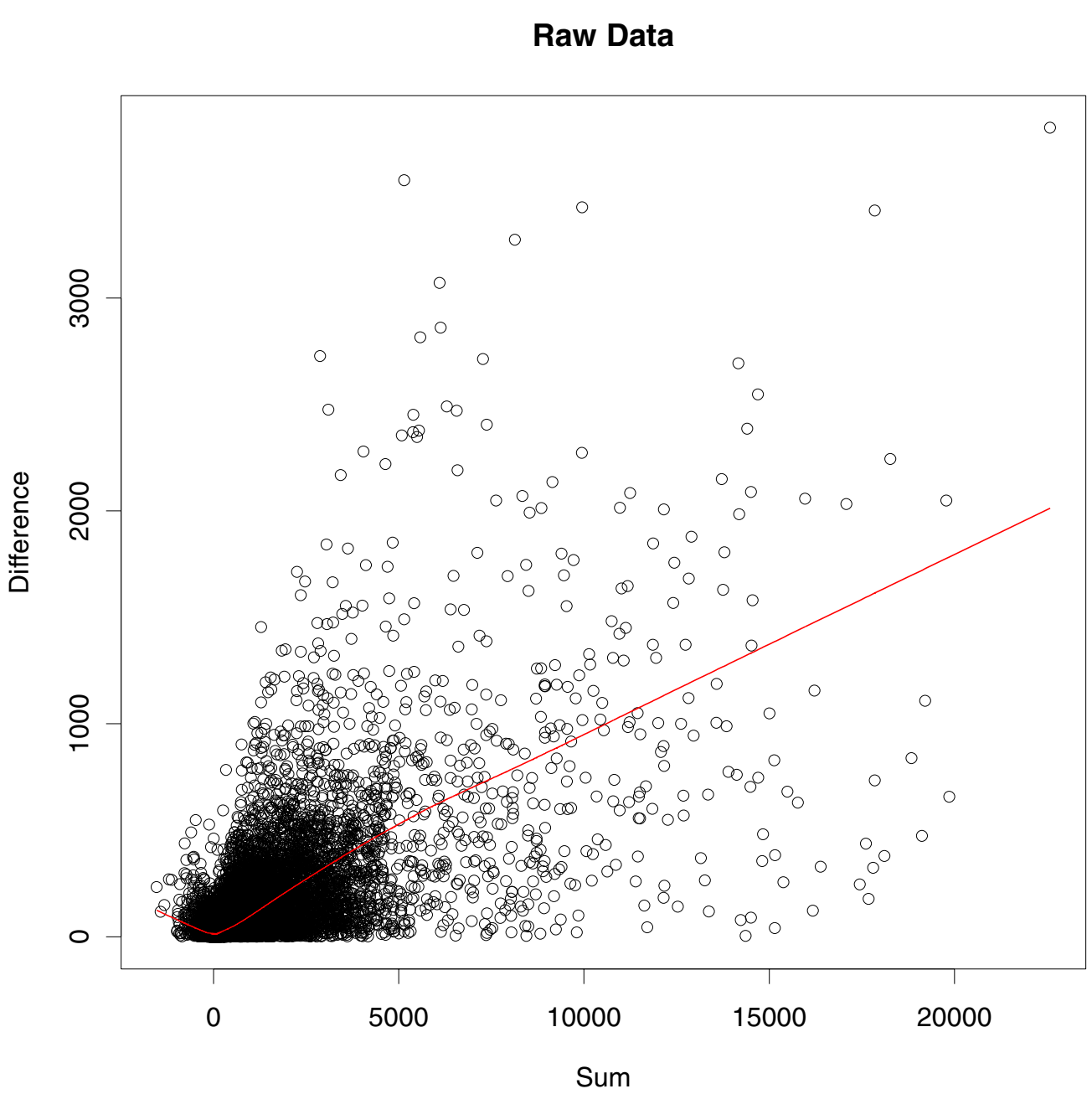
- Adjust p-values for multiplicity using, for example, the False Discovery Rate method.

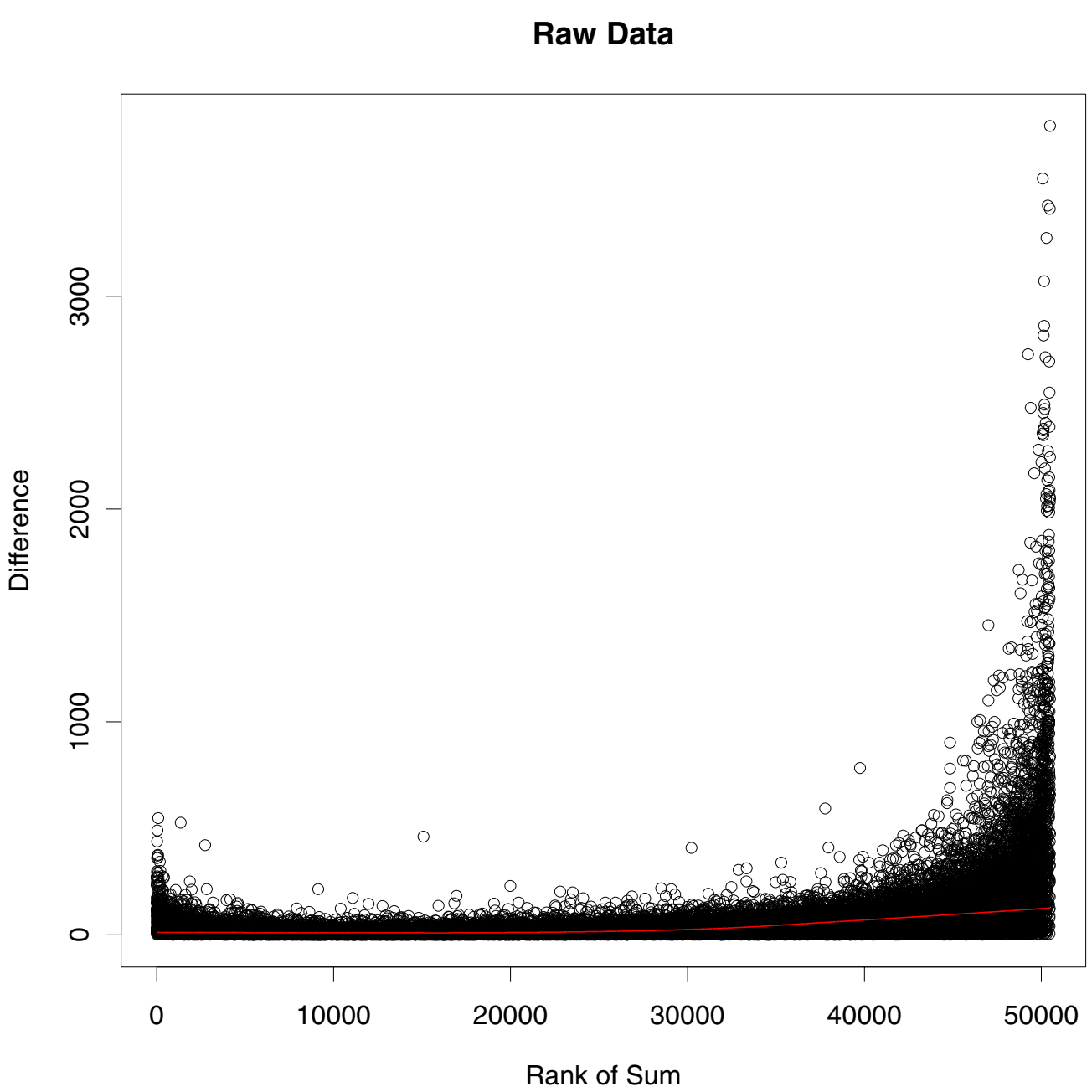
- Provide list of differentially expressed genes

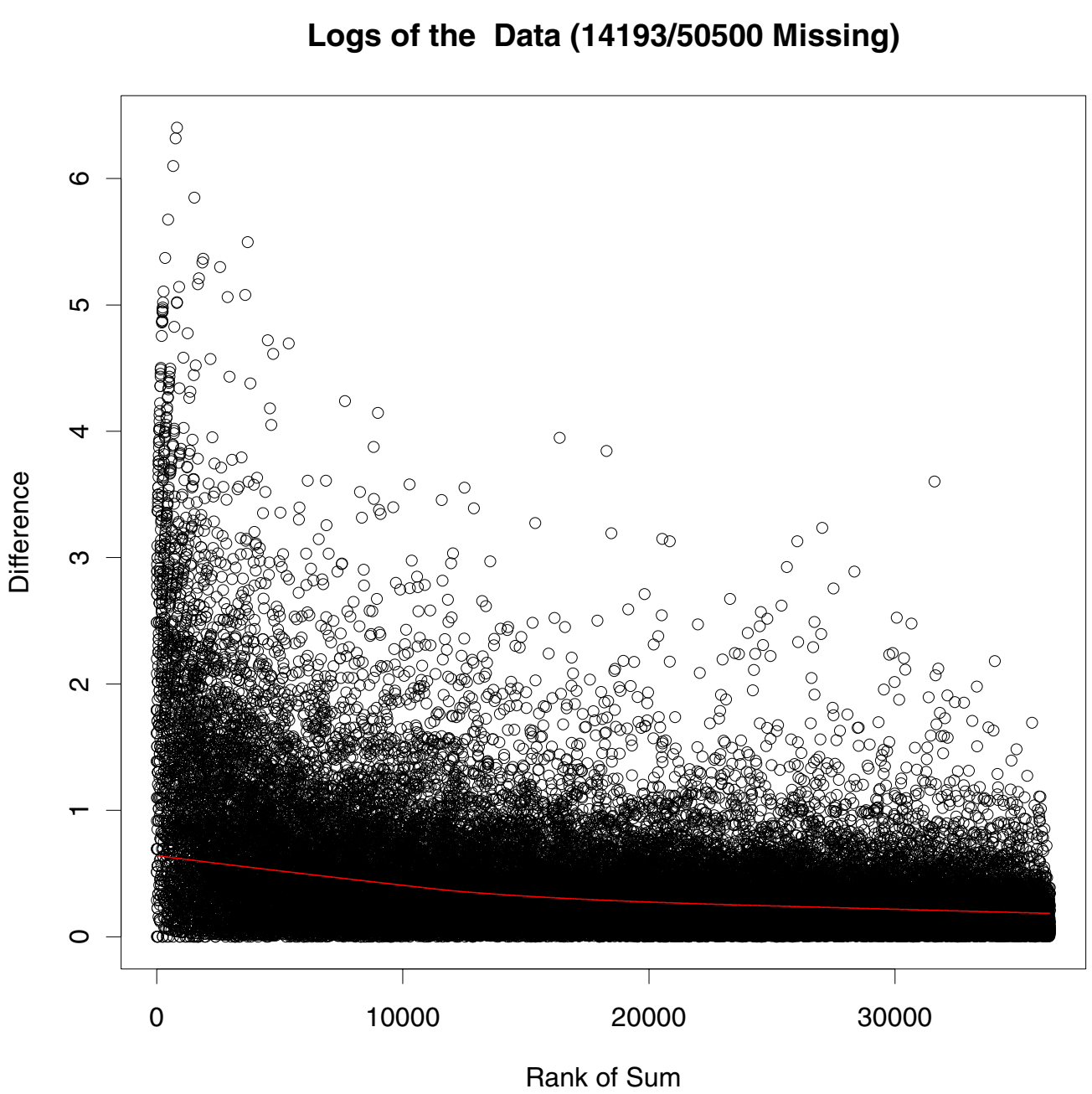
- Investigate identified genes statistically and by biological follow-up experiments.

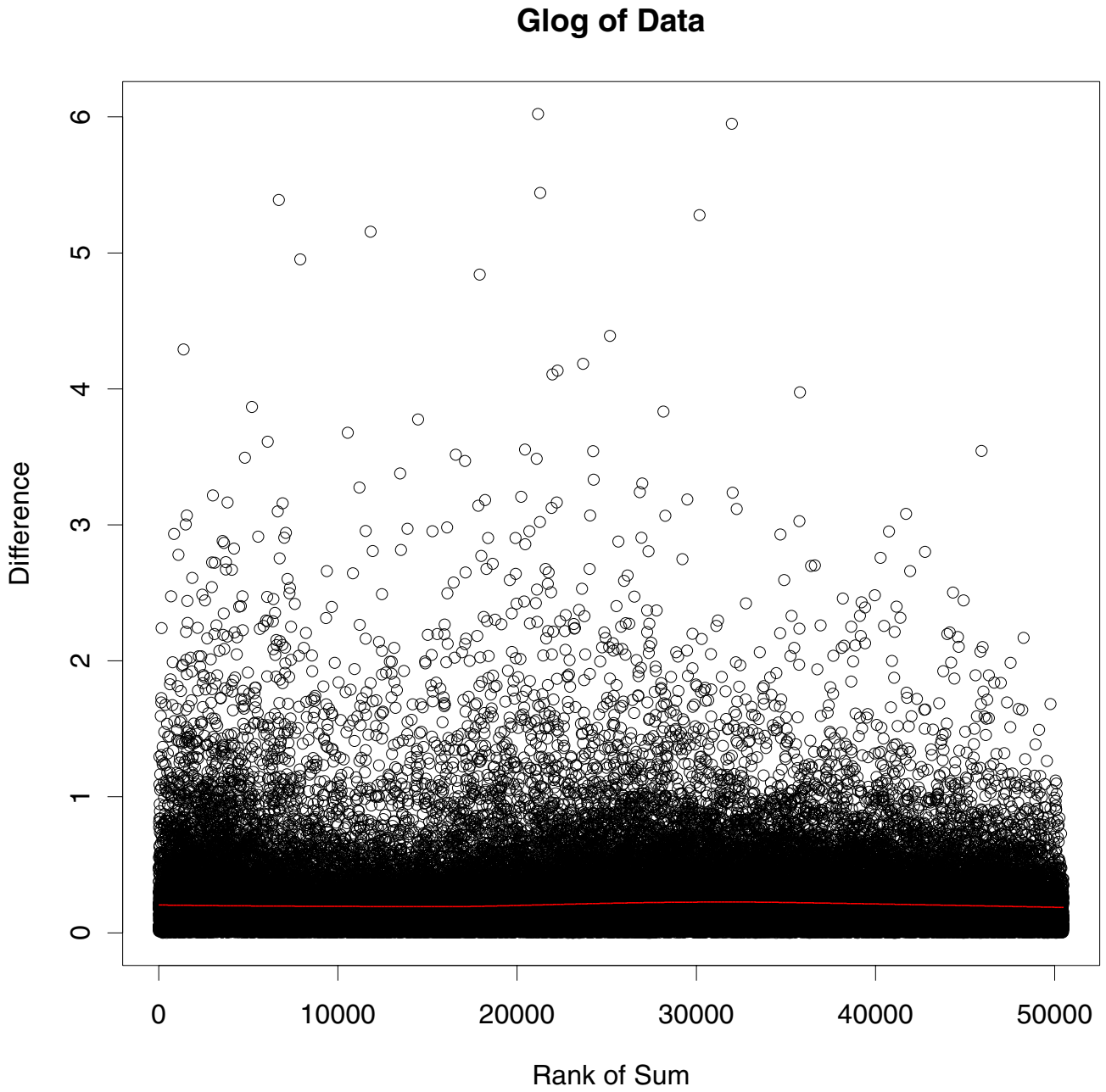
Structure of Example Data

Gene	Group 1		Group 2		Group 3		Group 4	
ID	1	2	3	4	5	6	7	8
1	y_{111}	y_{112}	y_{123}	y_{124}	y_{135}	y_{136}	y_{147}	y_{148}
2	y_{211}	y_{212}	y_{223}	y_{224}	y_{235}	y_{236}	y_{247}	y_{248}
3	y_{311}	y_{312}	y_{323}	y_{324}	y_{335}	y_{336}	y_{347}	y_{348}
4	y_{411}	y_{412}	y_{423}	y_{424}	y_{435}	y_{436}	y_{447}	y_{448}
5	y_{511}	y_{512}	y_{523}	y_{524}	y_{535}	y_{536}	y_{547}	y_{548}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮


Raw Data
Difference
0
1000
2000
3000
0
5000
10000
15000
20000
Sum


Raw Data
Difference
0
1000
2000
3000
0
10000
20000
30000
40000
50000
Rank of Sum


Logs of the Data (14193/50500 Missing)
Difference
0
1
2
3
4
5
6
0
10000
20000
30000
Rank of Sum


Glog of Data
Difference
0
1
2
3
4
5
6
0
10000
20000
30000
40000
50000
Rank of Sum

The model we use is

$$h_{\lambda,\alpha}(y_{ijk}) = \mu_i + n_k + \beta_{ij} + \epsilon_{ijk}$$

- We estimate all the parameters by normal maximum likelihood after the fashion of Box and Cox.

- The likelihood for the data y will be the same as the MSE from the model for the transformed Jacobian-corrected data z.

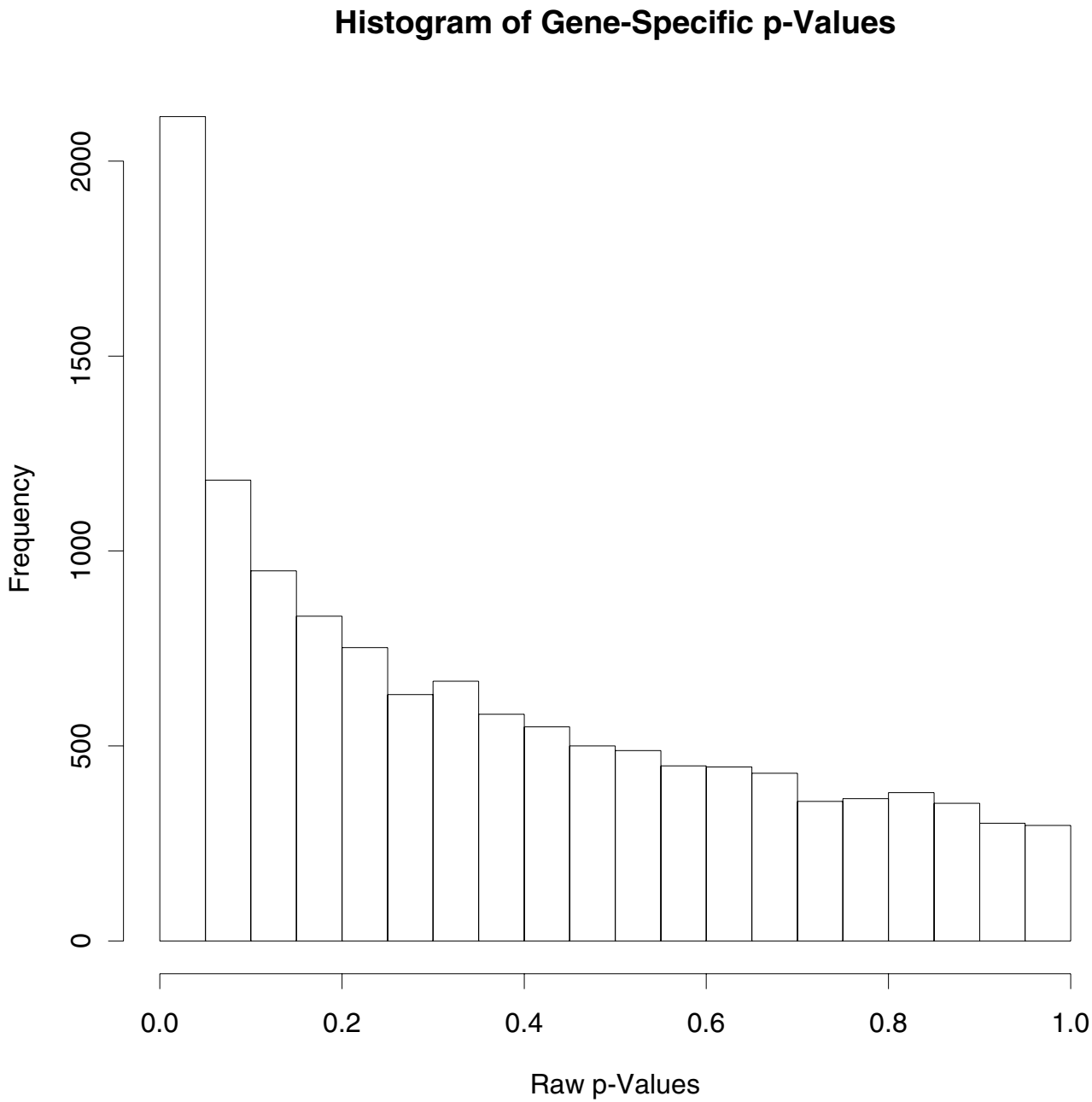
We can determine the MSE by transforming the data, multiplying each observation by the Jacobian, which is $\text{gm}(\sqrt{(y-\alpha)^2+\lambda})$, normalizing the chips additively, and fitting the model to each gene separately, then summing the individual MSE's for each gene. This works even with missing data, with only the normalization constants being slightly compromised.

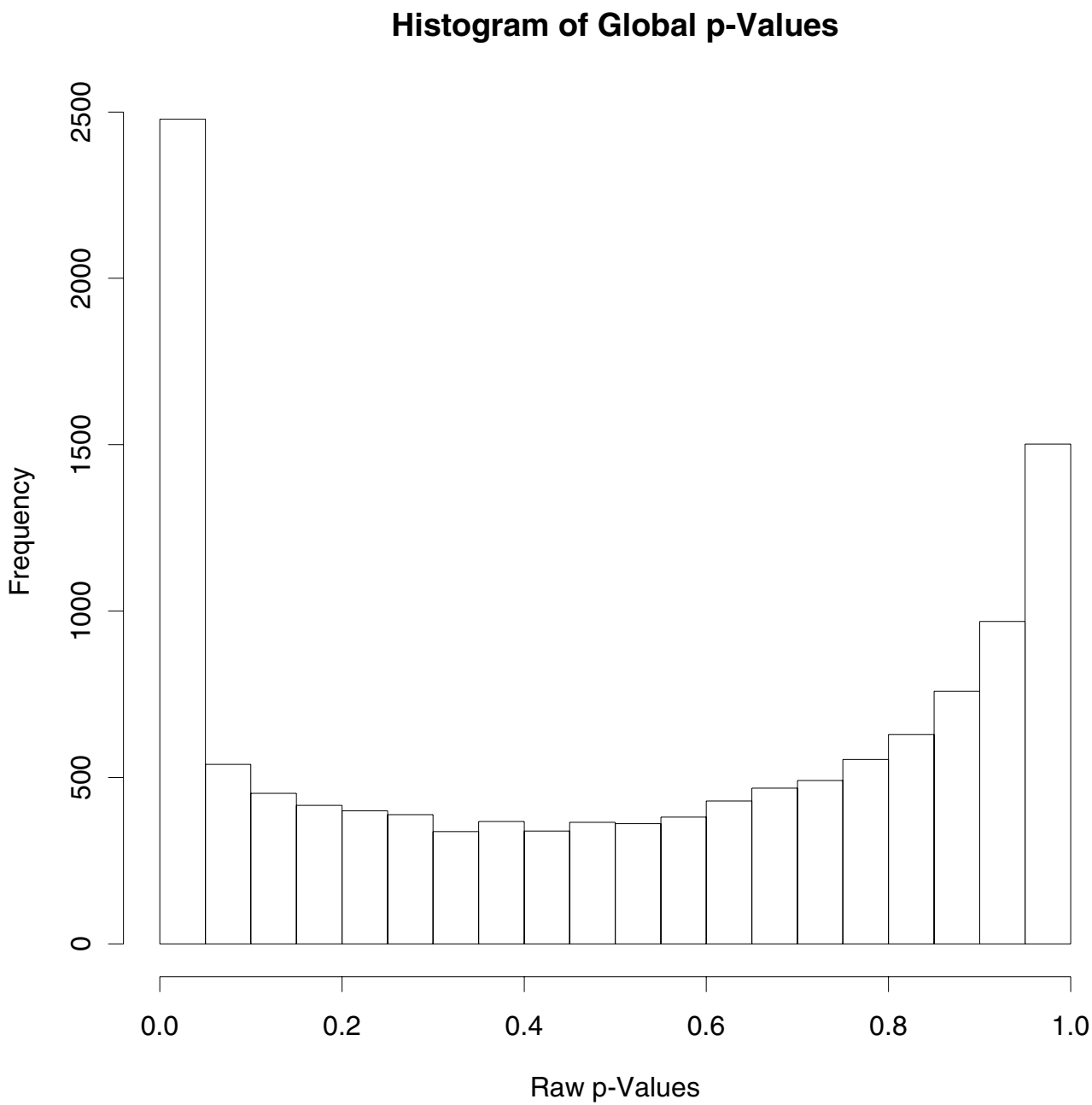
We can't fit the model for the full data set using most linear model software. The X matrix has $12{,}625 \times 8 = 101{,}000$ rows and $12{,}625 \times 4 + 7 = 50{,}507$ columns. The $X^\top X$ matrix is then 50,507 by 50,507, containing 2.55×10^9 8-byte reals, or 19GB!

- We can test for differential expression for a given gene by analyzing the transformed, normalized data in a standard one-way AVOVA.

- We can use as a denominator the gene-specific 4df MSE from that ANOVA. This is valid but not powerful.

- We can use the overall 50,493df MSE as a denominator. This is powerful, but risky.


Histogram of Gene-Specific p-Values
Frequency
0
500
1000
1500
2000
0.0
0.2
0.4
0.6
0.8
1.0
Raw p-Values


Histogram of Global p-Values
Frequency
0
500
1000
1500
2000
2500
0.0
0.2
0.4
0.6
0.8
1.0
Raw p-Values

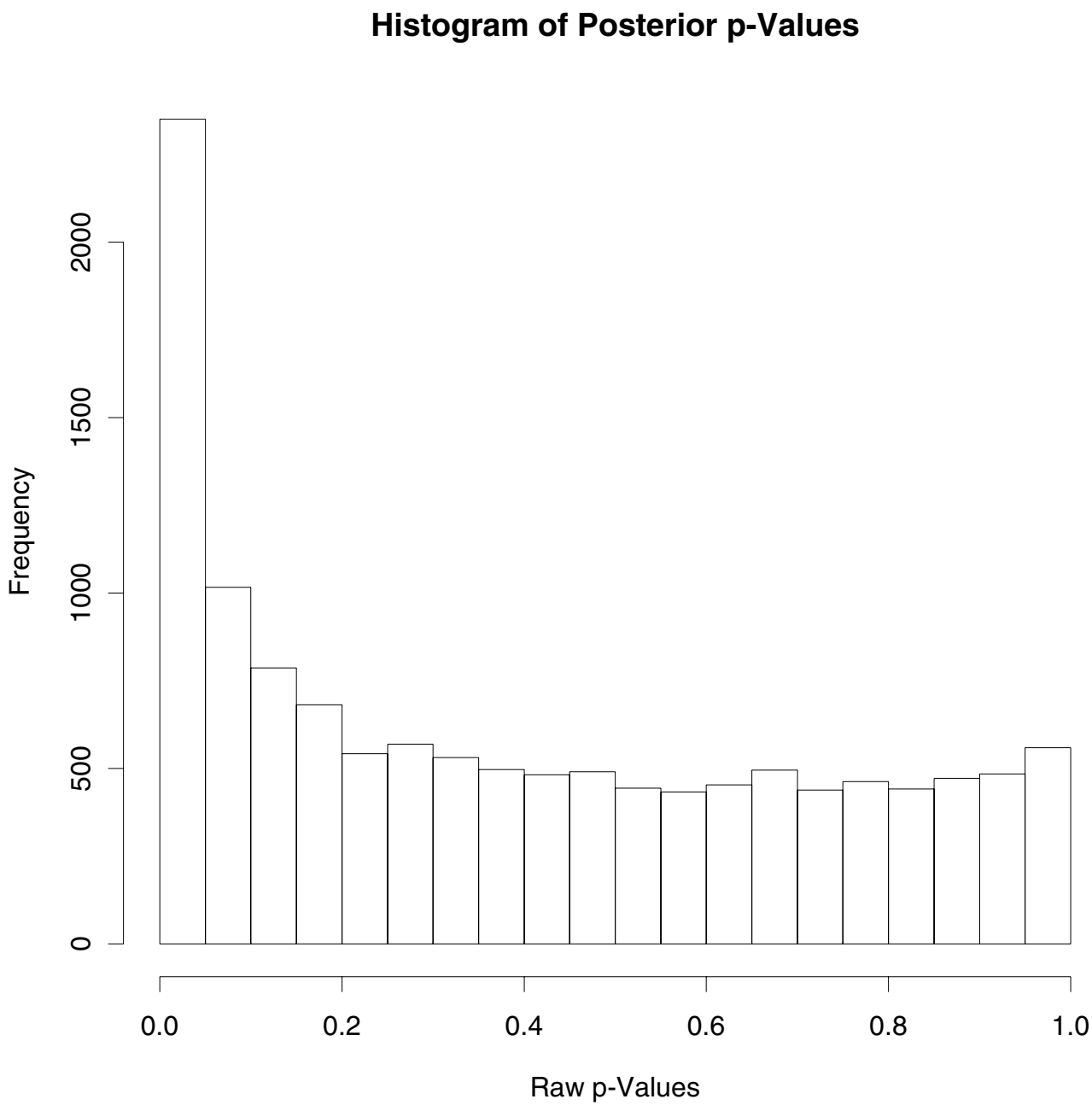
- As an alternative, we can postulate a hierarchical model in which the gene-specific true variances are generated from an inverse gamma, the conjugate prior under normality.

- The overall MSE is 0.102. The variance across genes of the 4df estimates under homogeneity should be approximately $2\sigma^4/4 = (.102)^2/2 = 0.0057$. Instead, the variance is 0.0556, which is 10 times larger.

- We fit an inverse gamma to the mean and variance of the gene-specific variances in the hierarchical model, yielding $\alpha = 2.308$, $\beta = 7.520$, $\nu = 2\alpha = 4.615$.

- We have $1/\alpha\beta = .05763$, which will be used as the prior best estimate. Thus number is the reciprocal mean prior precision. The empirically estimated degrees of freedom for the inverse gamma prior is 4.6.

- The posterior best estimate MSE is a weighted average of the gene-specific MSE (with weight 4/8.6) and the prior best estimate (with weight 4.6/8.6) and has 8.6 degrees of freedom. Like the prior, it is of the form $1/\alpha\beta$, for posterior estimates of α and β.


Histogram of Posterior p-Values
Frequency
0
500
1000
1500
2000
0.0
0.2
0.4
0.6
0.8
1.0
Raw p-Values

"5% Significant" Genes by Several Methods

MSE Source	TWER	FWER	FDR
Gene-Specific	2114	1	18
Global	2478	571	1516
Posterior	2350	29	508

Probe-Level Analysis of Affymetrix Data

- The previous analysis started with MAS 4.0 average difference (for the sake of convenience).

- Likely, it is better to transform the probe level data and determine the gene expression value from the PM probes only.

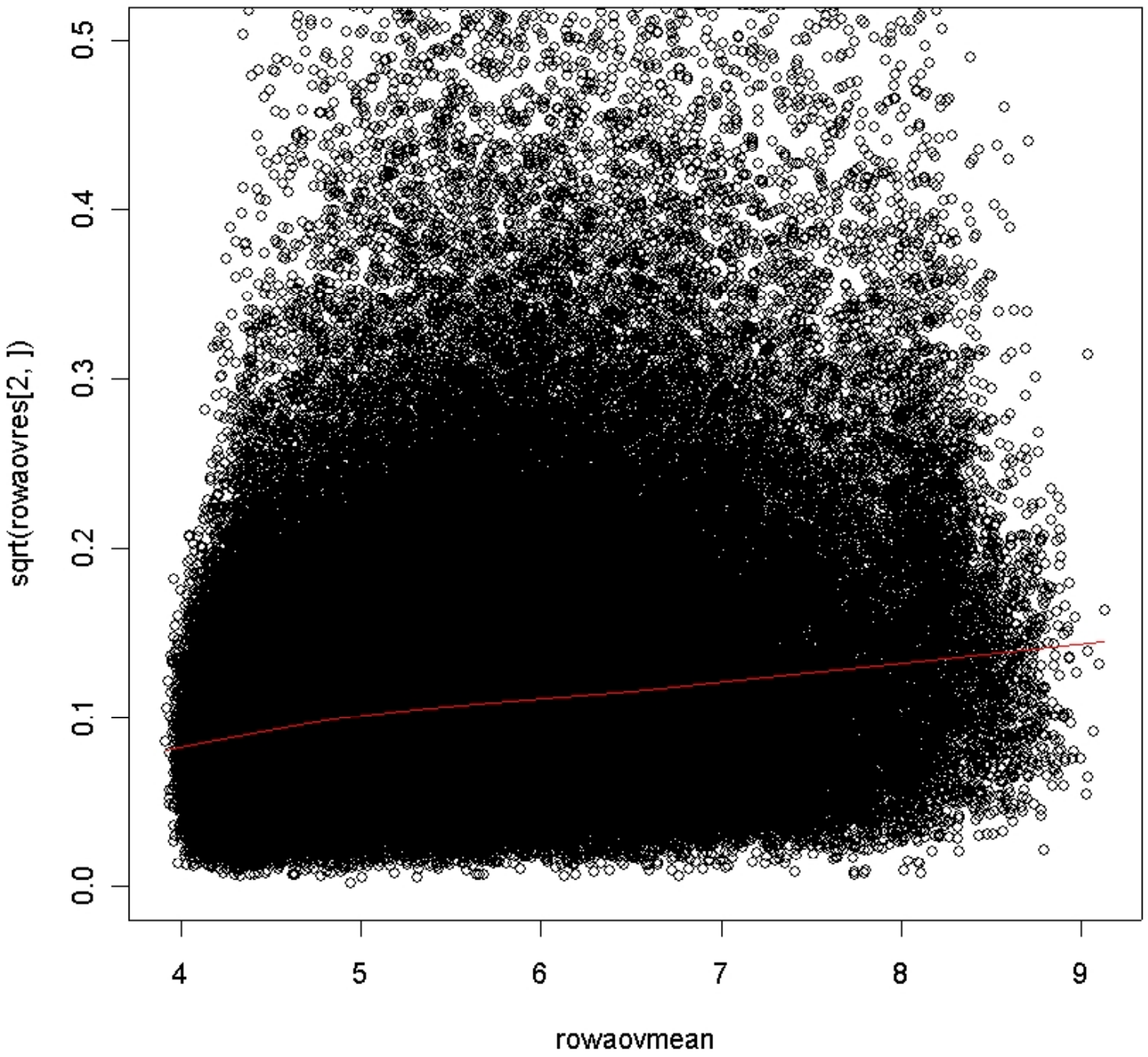
- This is already being done with various transformation and normalization procedures by a number of investigators.

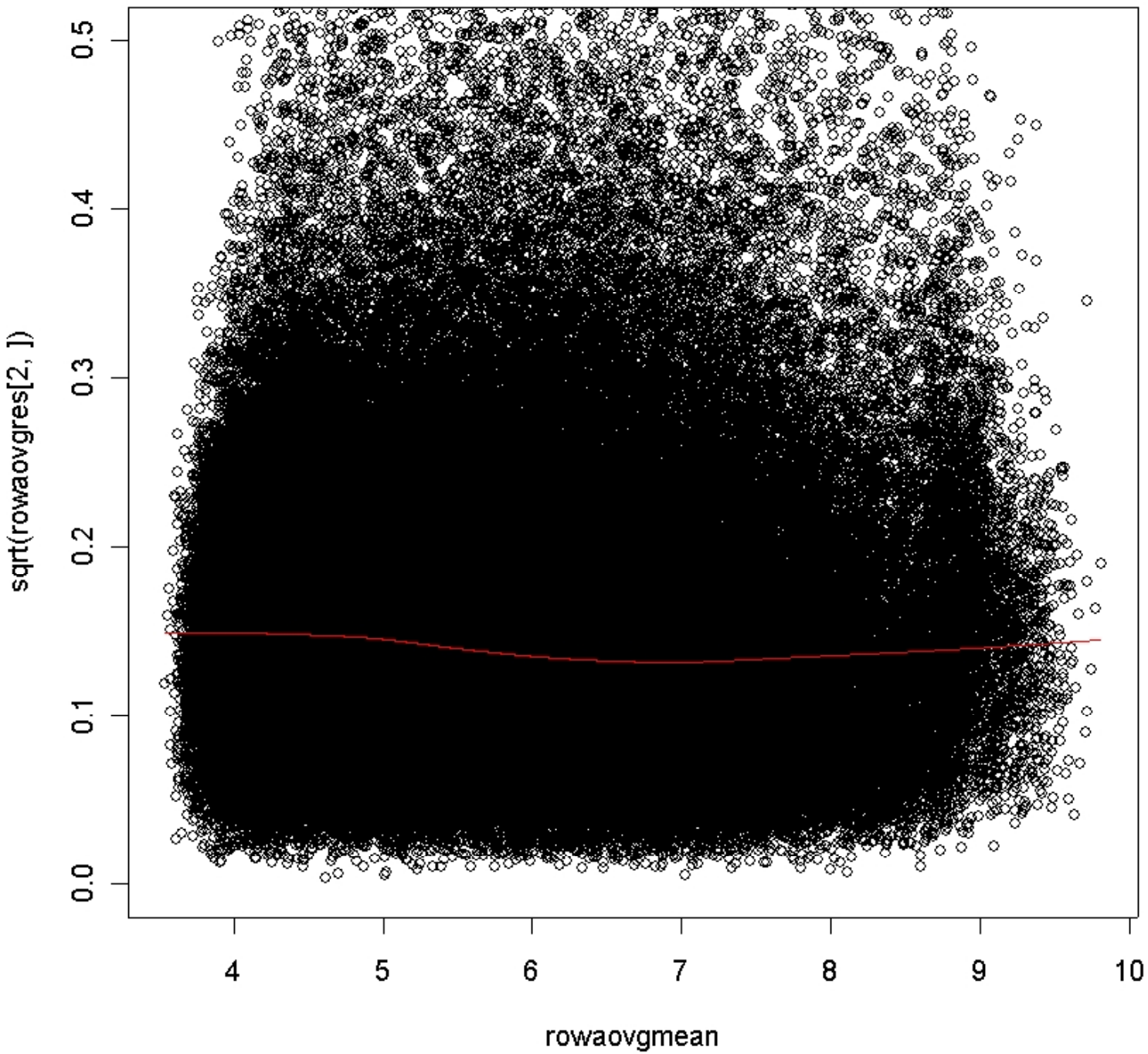
Which Transformation?

- We can use the log of the PM probes, the glog, the started log, or many others.

- Definitive evidence will use dilution and spike-in data as well as many other data sets for comparison.

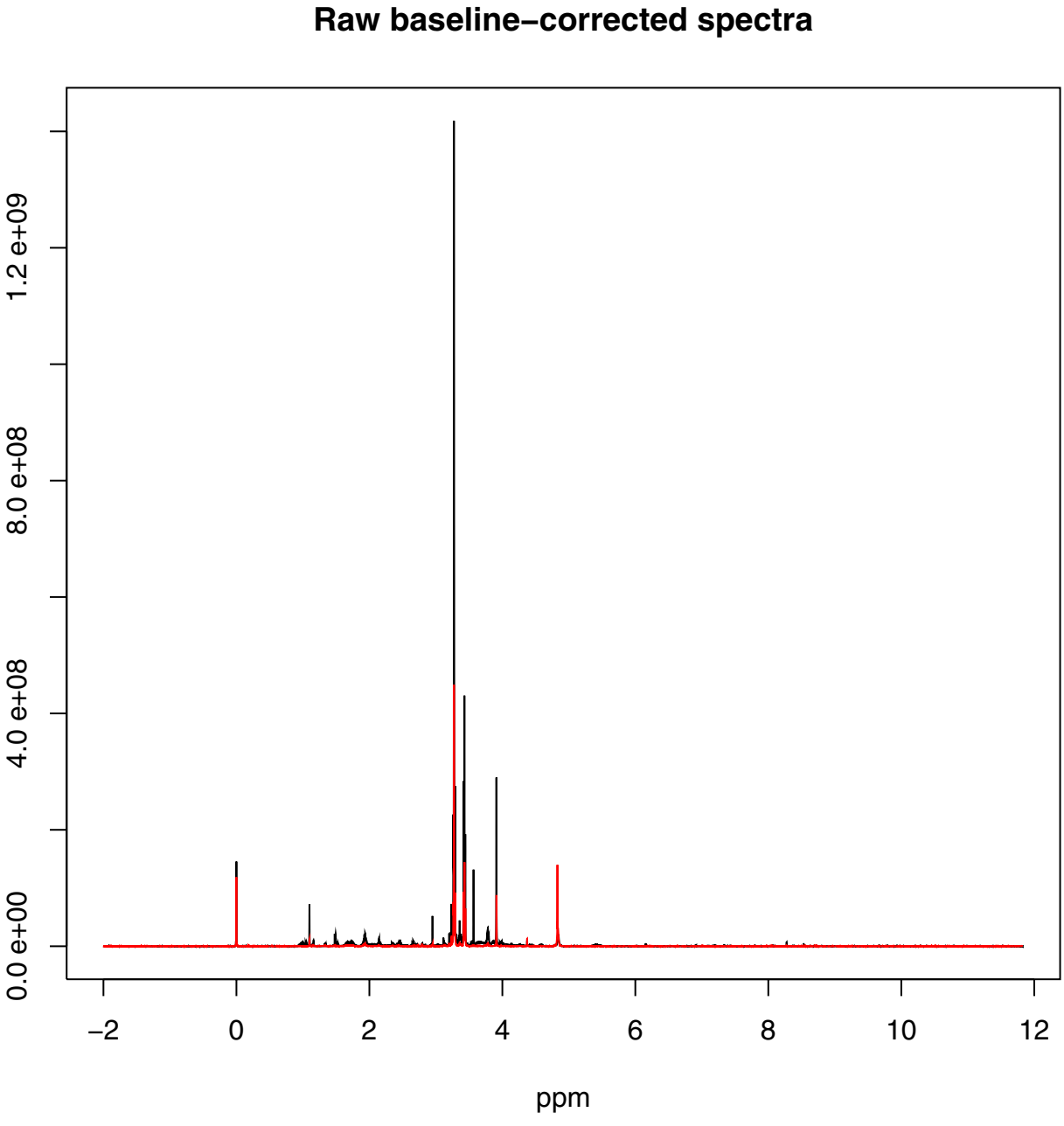
- One desirable property is stability of variance.

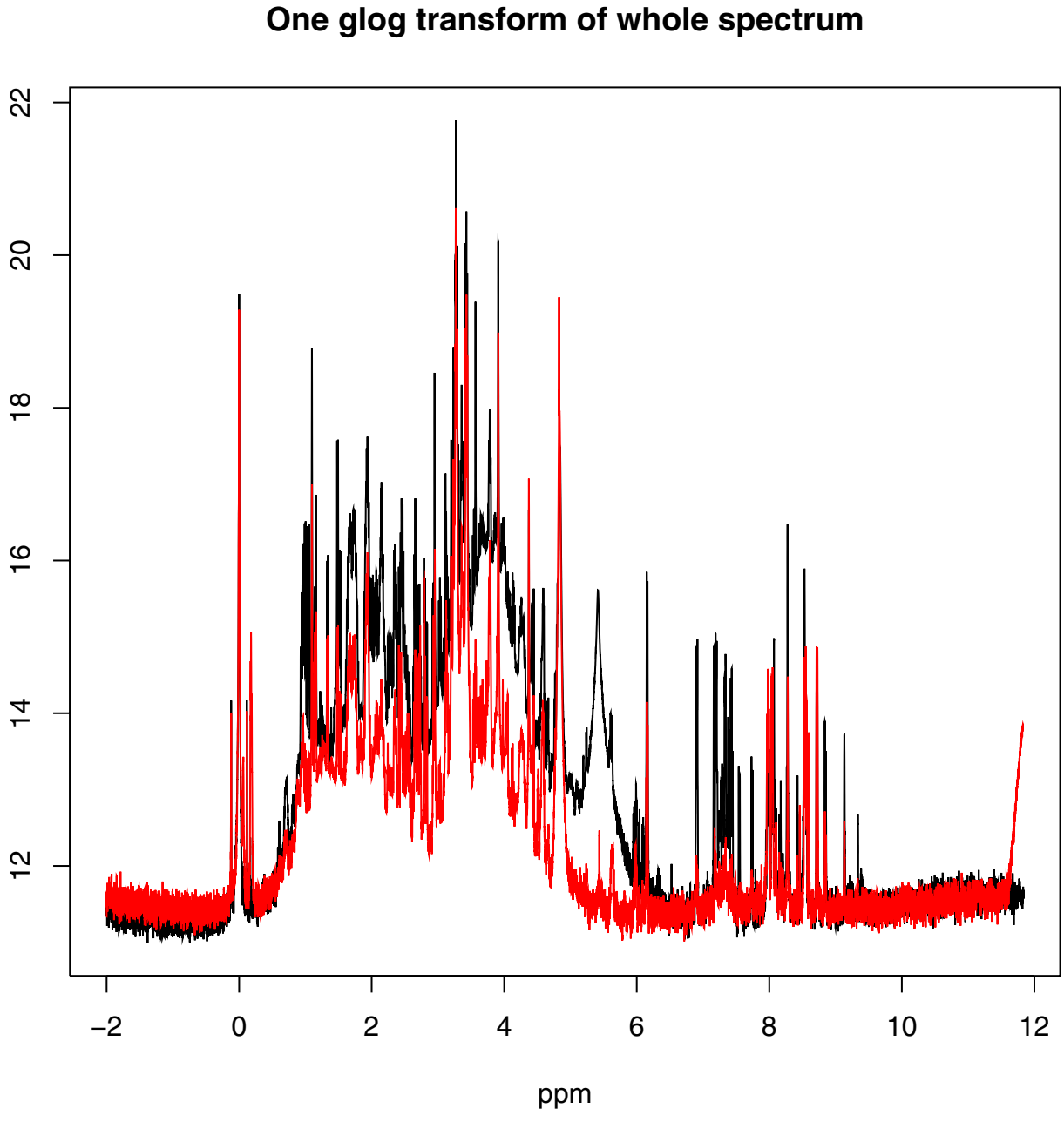

sqrt(rowaovres[2,])
0.0
0.1
0.2
0.3
0.4
0.5
4
5
6
7
8
9
rowaovmean

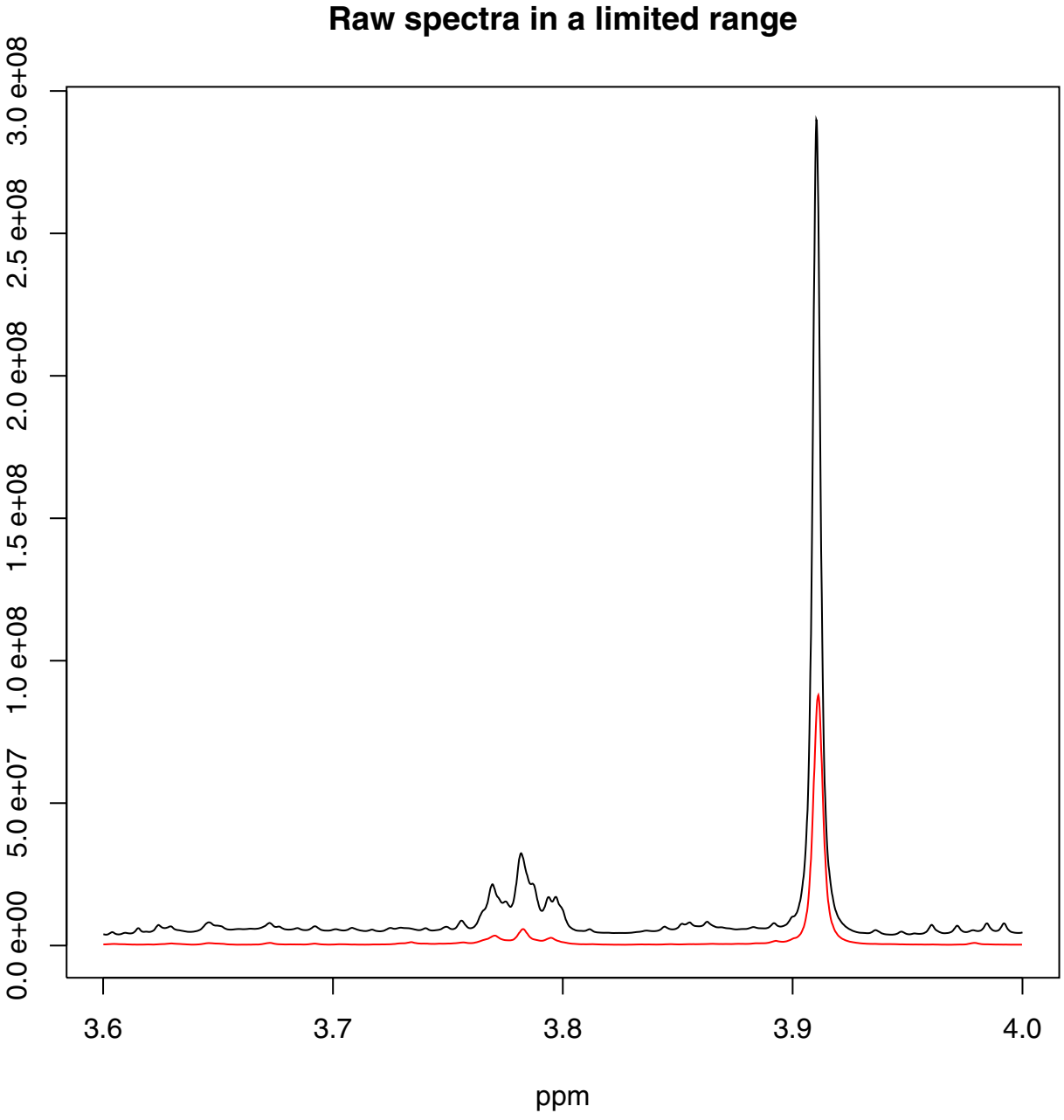

sqrt(rowaovgres[2,])
0.0
0.1
0.2
0.3
0.4
0.5
4
5
6
7
8
9
10
rowaovgmean

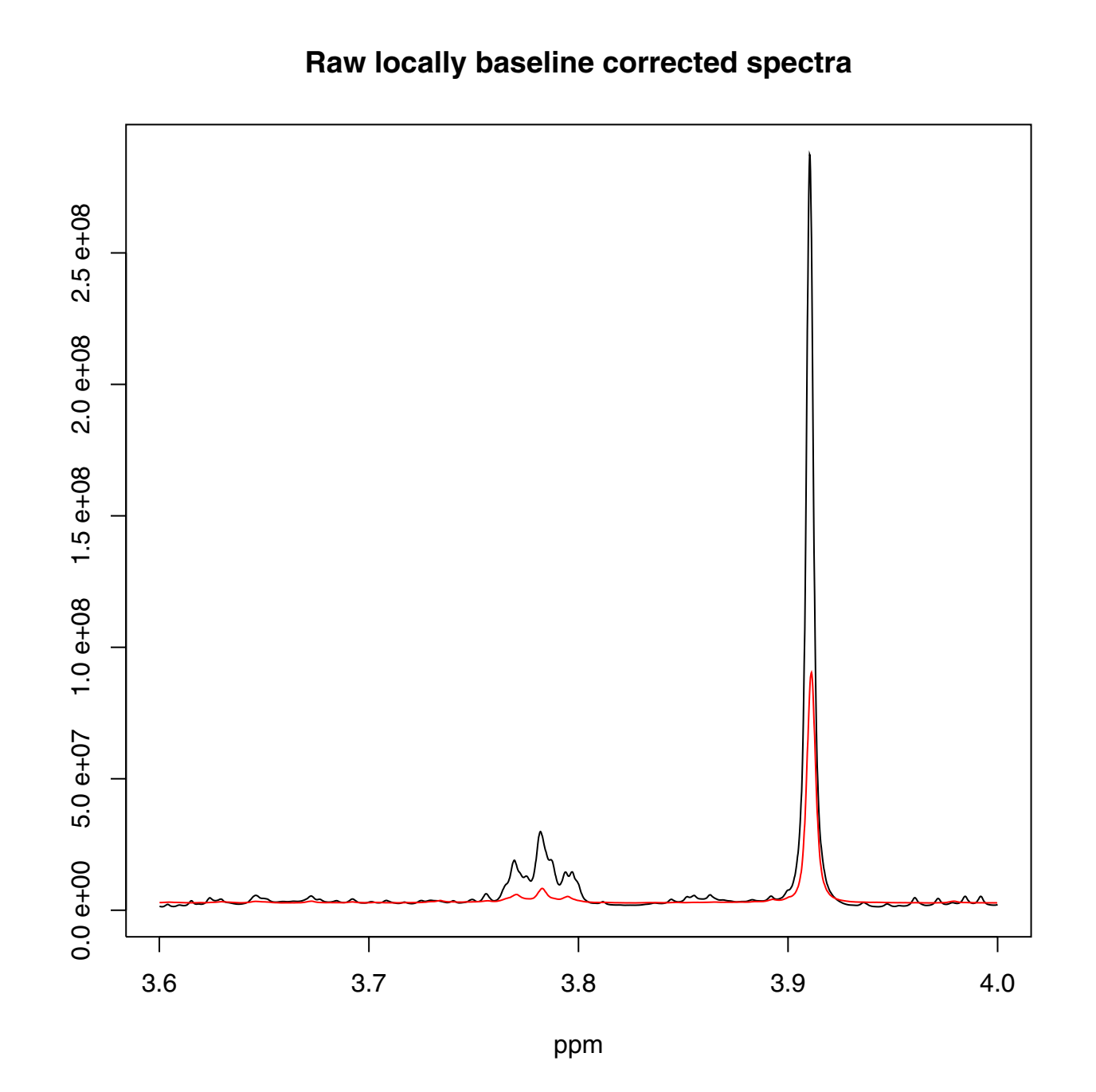
Metabolomics by NMR Spectroscopy

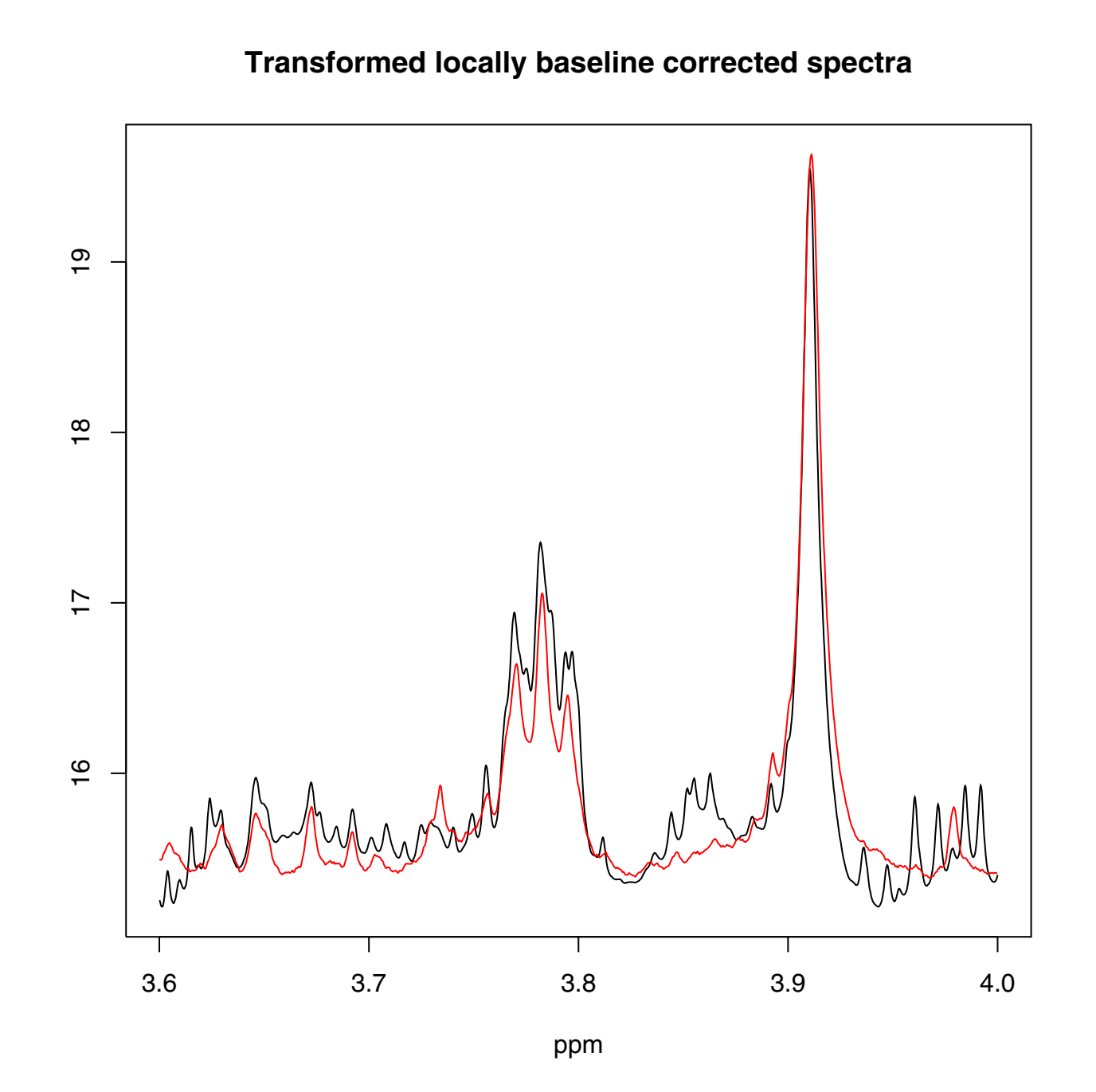
- Spectra need to be baseline corrected.
- After baseline correction, the peaks are of widely varying magnitudes, and some of the data are negative.
- The glog is a plausible transformation to help in the analysis of these data also.

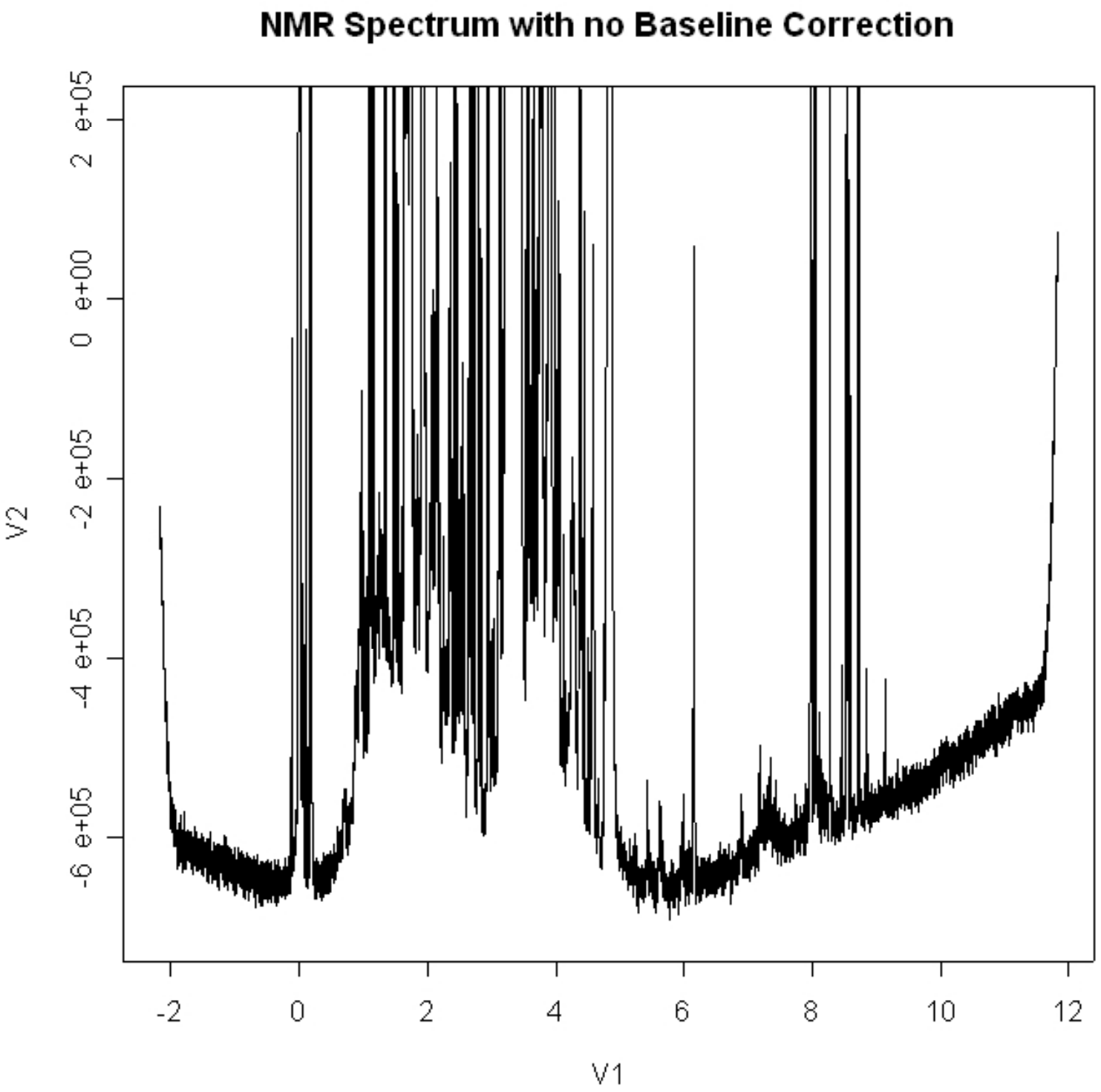

Raw baseline−corrected spectra
0.0 e+00
4.0 e+08
8.0 e+08
1.2 e+09
−2
0
2
4
6
8
10
12
ppm


One glog transform of whole spectrum
22
20
18
16
14
12
−2
0
2
4
6
8
10
12
ppm


Raw spectra in a limited range
0.0 e+00
5.0 e+07
1.0 e+08
1.5 e+08
2.0 e+08
2.5 e+08
3.0 e+08
3.6
3.7
3.8
3.9
4.0
ppm


Raw locally baseline corrected spectra
0.0 e+00
5.0 e+07
1.0 e+08
1.5 e+08
2.0 e+08
2.5 e+08
3.6
3.7
3.8
3.9
4.0
ppm


Transformed locally baseline corrected spectra
19
18
17
16
3.6
3.7
3.8
3.9
4.0
ppm


NMR Spectrum with no Baseline Correction
V2
2 e+05
0 e+00
-2 e+05
-4 e+05
-6 e+05
-2
0
2
4
6
8
10
12
V1

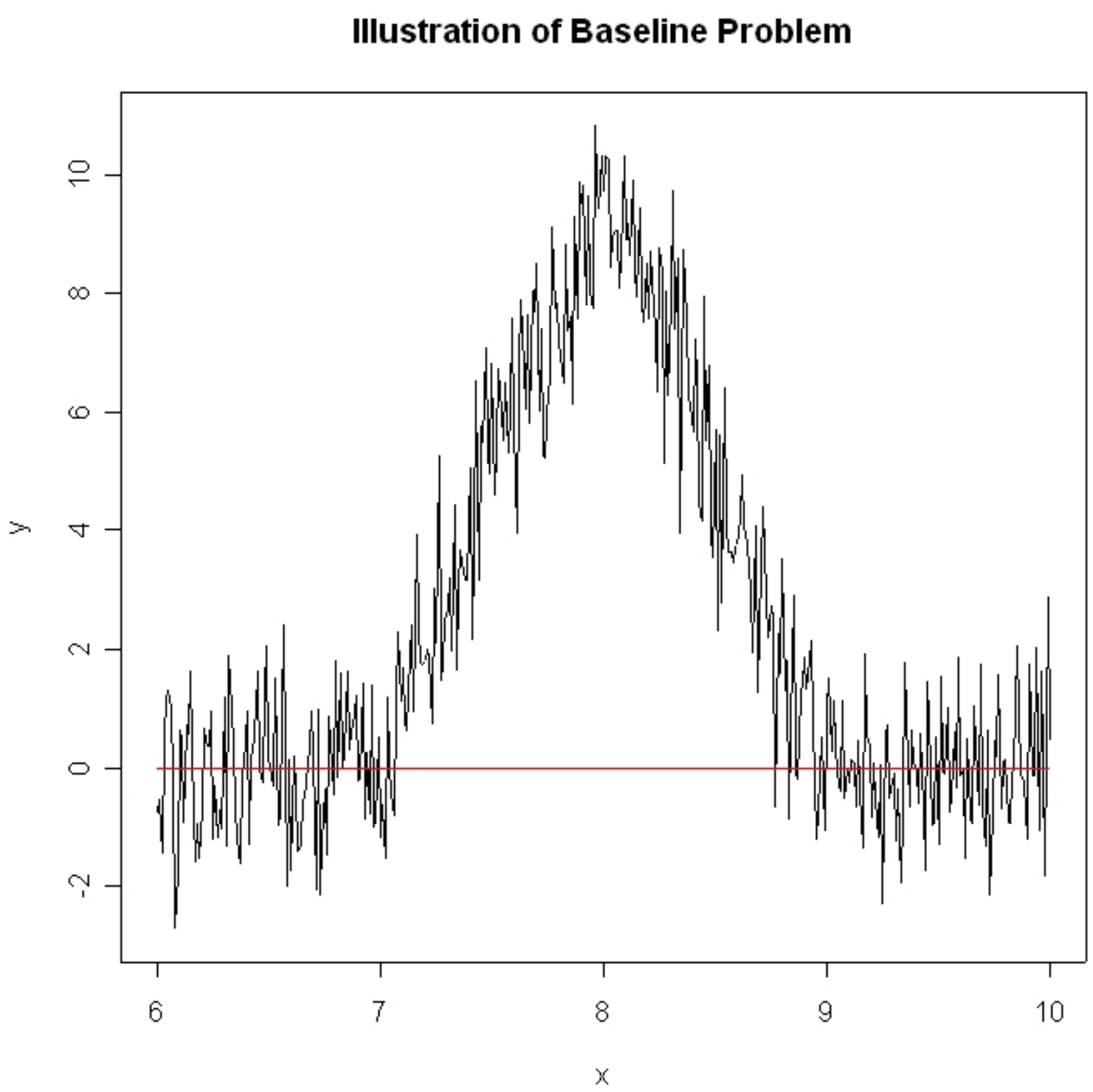


Illustration of Baseline Problem
y
x
10
8
6
4
2
0
-2
6
7
8
9
10

Conclusion

- Gene expression, proteomics, and metabolomics data present many interesting statistical challenges.

- We have presented a model for measurement error that guides the transformation of the data, helps determine significance of changes, and allows more sophisticated analysis.

- The two-component model seems to fit microarray and other assay data well.

- A properly chosen transformation can stabilize the variance and improve the statistical properties of analyses.

- Slide normalization and analysis of two-color arrays is made easier by this transformation.

- Other statistical calculations such as the analysis of variance that assume constant variance are also improved.

- After removal of systematic dependence of the variance on the mean, the remaining sporadic variation in the variance can be accounted for by a simple empirical Bayes method.

- We are now working to apply these methods to other types of data such as MALDI-TOF proteomics, LC/MS and LC/TOF metabolomics, NMR spectroscopy metabolomics, and GC lipid metabolomics. The variables measured are a large number of peak heights or areas, or a large number of binned spectroscopic values

- Papers are available at **www.cipic.ucdavis.edu/∼dmrocke** or by mail and e-mail.

Acknowledgements

Lab Personnel: Collaborators and Students

Faculty

Sue Geller (Texas A&M)

David Woodruff

Research Staff

Jian Dai

Parul Purohit

Students and Former Students

Blythe Durbin

Johanna Hardin (Pomona College)

Jinjin Liang

Danh Nguyen (Texas A&M)

Machelle Wilson (University of Georgia)

Lei Zhou

UC Davis Collaborators

Matt Bartosiewicz

Alan Buckpitt

Jeff Gregg

Paul Hagerman

Bruce Hammock

Mark Viant

Outside Collaborators

Doug Hawkins (University of Minnesota)

Steve Watkins (Lipomics, Inc.)

Funding

NSF

NIEHS

US EPA

UC Davis MIND Institute