Variability and Data Transformation for Gene Expression, Proteomics, and Metabolomics Data

David M. Rocke

Division of Biostatistics, School of Medicine,

Department of Applied Science, College of Engineering, &

Institute for Data Analysis and Visualization

University of California, Davis

Statistical Science for Genome Biology, BIRS
August, 2004

Overview

- Data from high-throughput biological assays such as gene expression arrays, proteomics by mass spectrometry, or metabolomics by NMR spectroscopy present many challenging problems for analysts.

- We present a variance model for these data that explains a number of problems currently facing users of these data.

- We present a class of data transformations specifically tuned to biological 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 replicates, from local properties of spectra, or using maximum likelihood in the context of a linear model analysis.

- We note that the remaining variance heterogeneity can often be accounted for with a hierarchical variance model that can be easily estimated in an empirical Bayes fashion (e.g., Wright and Simon 2003).

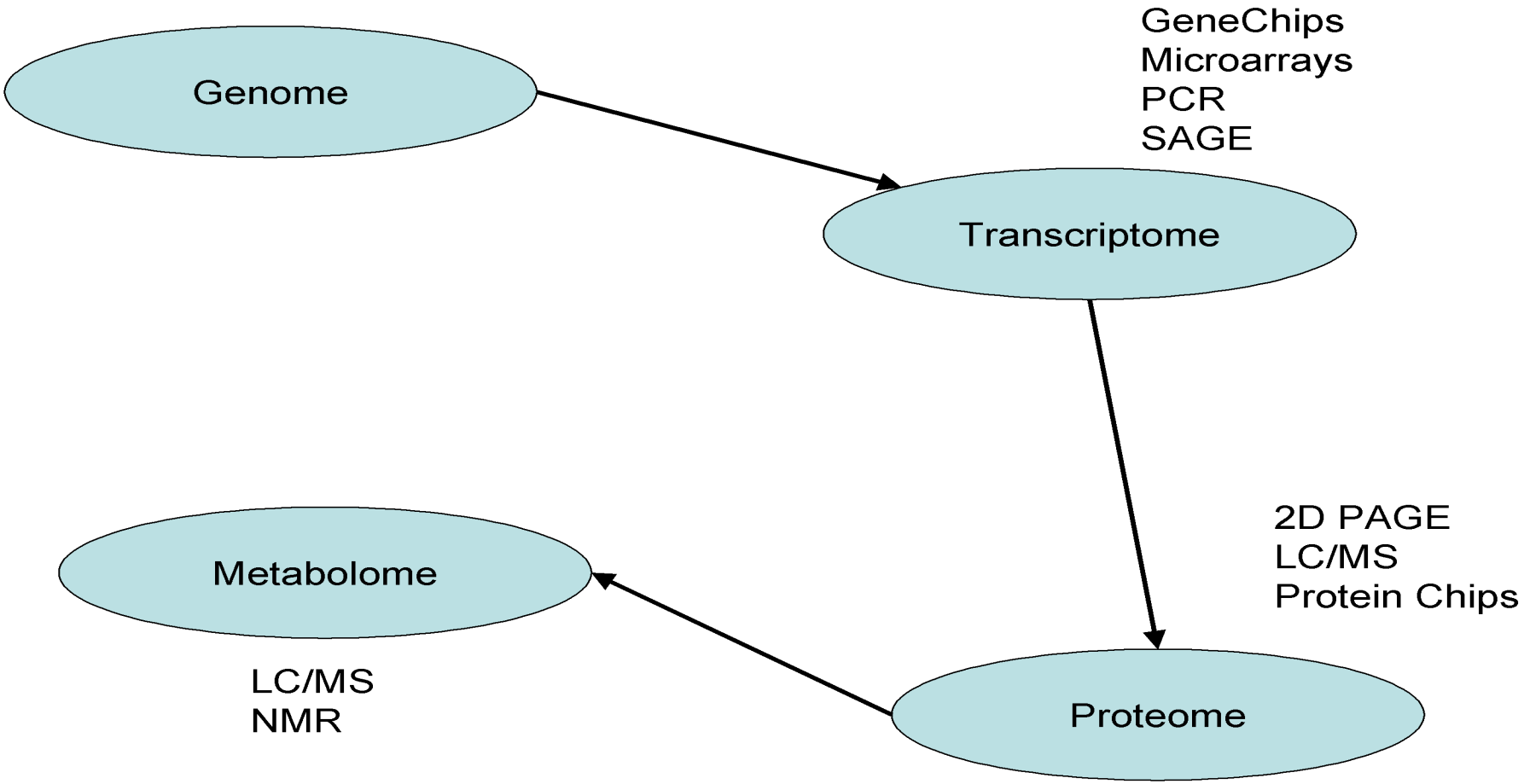
Omics Data

Genome Complement of all genes, or of all components of genetic material in the cell (mostly static).

Transcriptome Complement of all mRNA transcripts produced by a cell (dynamic).

Proteome Complement of all proteins in a cell, whether directly translated or produced by post-translational modification (dynamic).

Metabolome Complement of all metabolites other than proteins and mRNA; e.g., lipids, saccharides, etc (dynamic).


Genome
GeneChips
Microarrays
PCR
SAGE
Transcriptome
2D PAGE
LC/MS
Protein Chips
Metabolome
LC/MS
NMR
Proteome

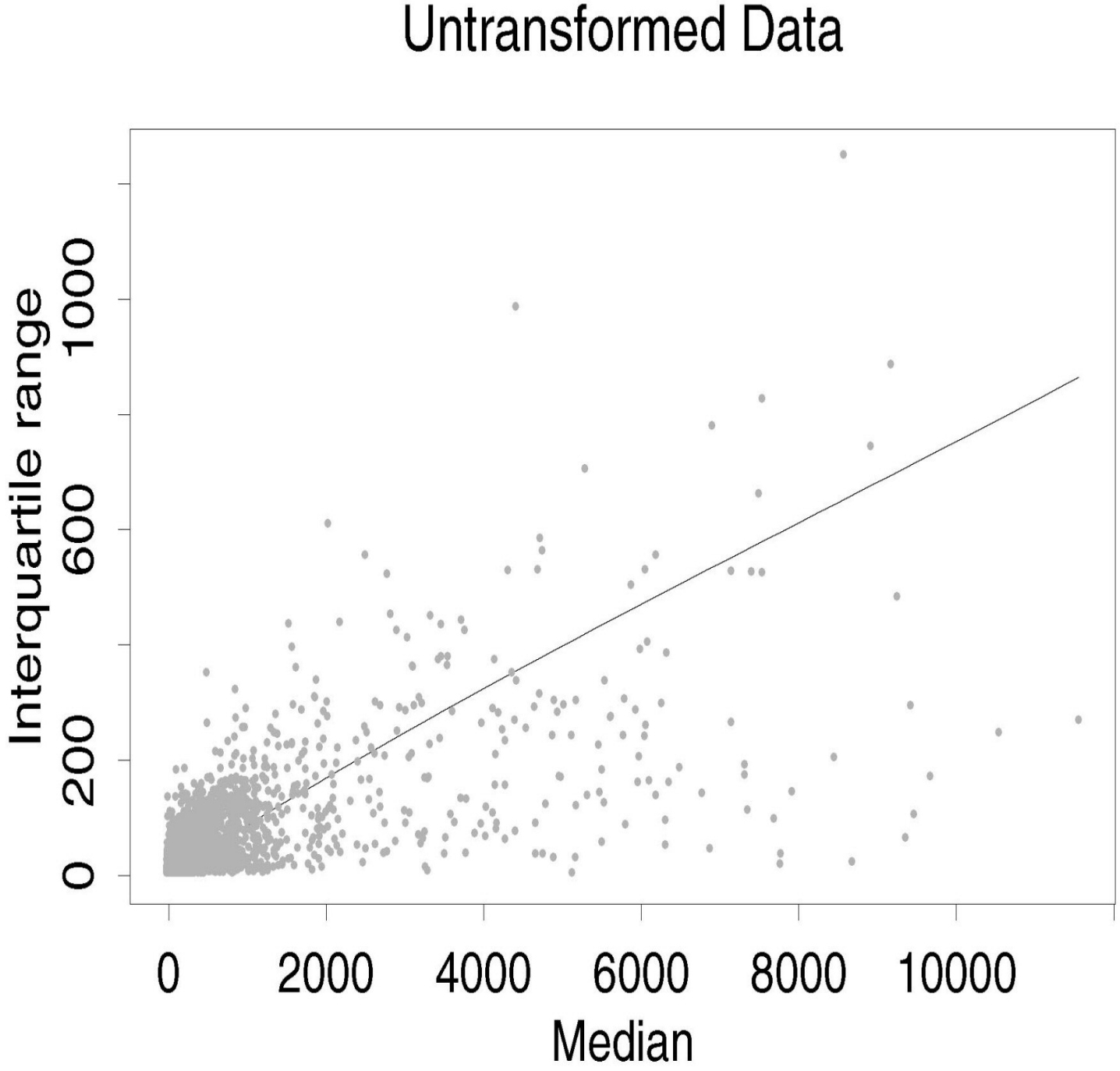
Common Variance Behavior

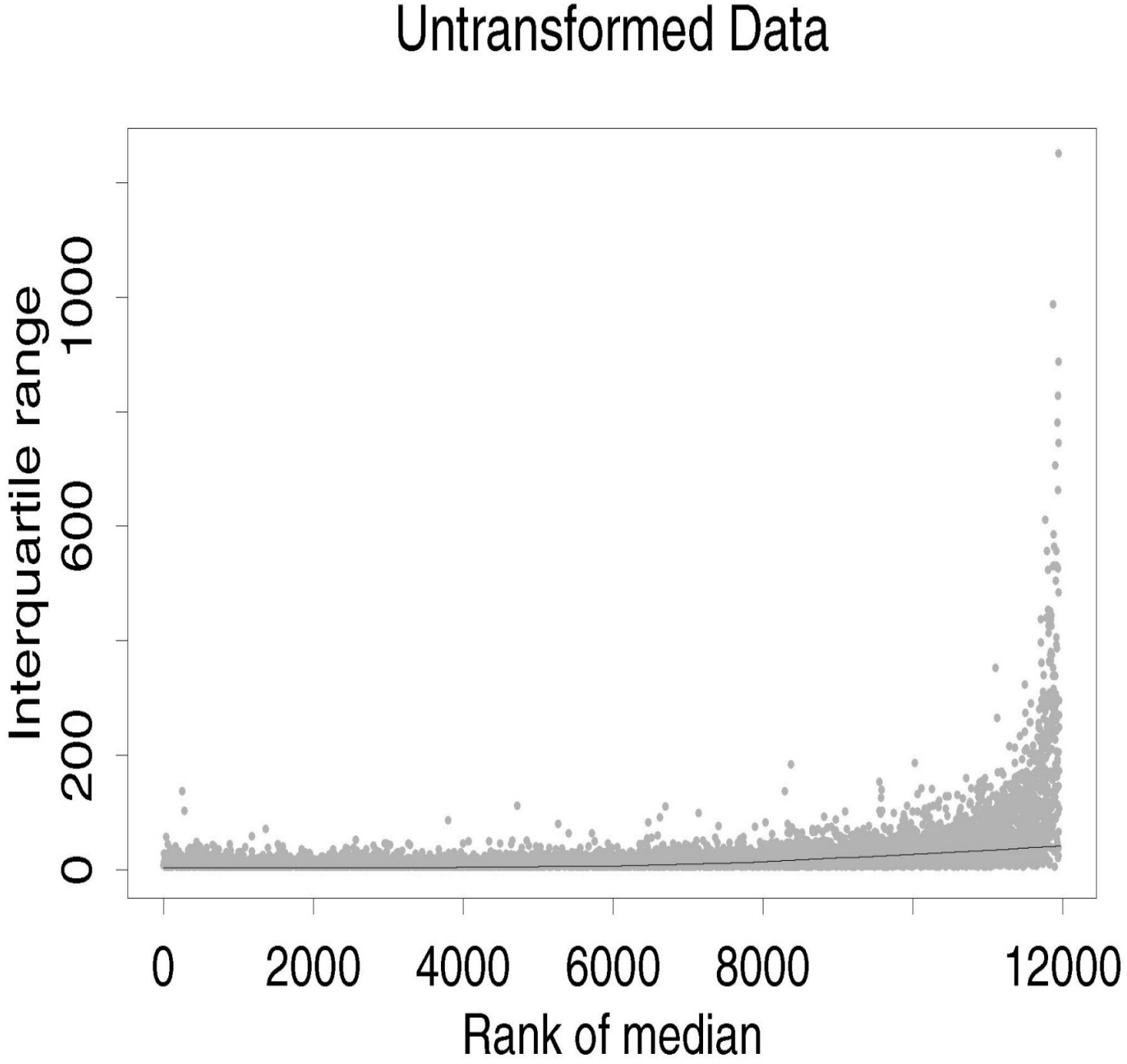
For many types of assays, the following are generally true:

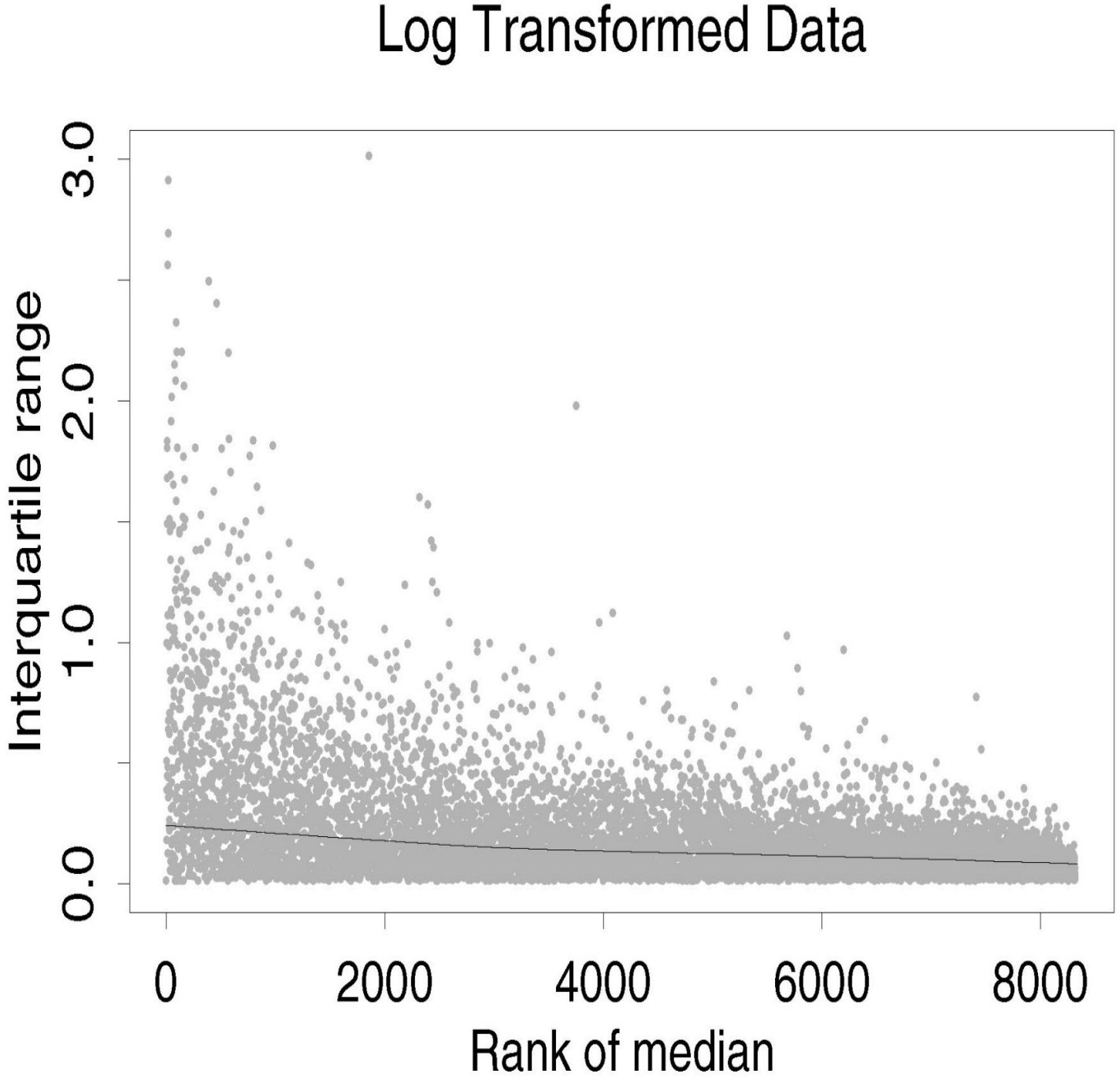
- For high concentrations, the standard deviation of the response is approximately proportional to the mean response, so that the CV is approximately constant.

- For low concentrations, the CV is much higher.

- Assay data are commonly analyzed on the log scale, so that for high levels the SD is approximately constant, but for low levels of the SD rises.


Untransformed Data
Interquartile range
0
200
600
1000
0
2000
4000
6000
8000
10000
Median


Untransformed Data
Interquartile range
0
200
600
1000
0
2000
4000
6000
8000
12000
Rank of median


Log Transformed Data
Interquartile range
0.0
1.0
2.0
3.0
0
2000
4000
6000
8000
Rank of median

- Comparisons of concentration are usually expressed as ratios or n-fold,, of which the logarithm would be well behaved, but only if both concentrations are well above zero.

- These phenomena occur in many measurement technologies, but are more important in high-throughput assays such as are often used for omics data.

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

- Which is biologically more important: an increase from 0 to 25 or an increase from 100 to 200?

Error Model for Assay Data

This model is appropriate for data in which it is not practical to form a calibration curve. A model for calibrated assays was analyzed in Rocke and Lorenzato (1995) and subsequent literature.

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 concentration level in arbitrary units, and α is the mean background/baseline (mean signal at zero concentration). Our best estimate of μ is $y - \widehat{\alpha}$, 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

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

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 ratios have constant variance only if both concentrations are well above zero.

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 *propagation of error* or the *delta method*.

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

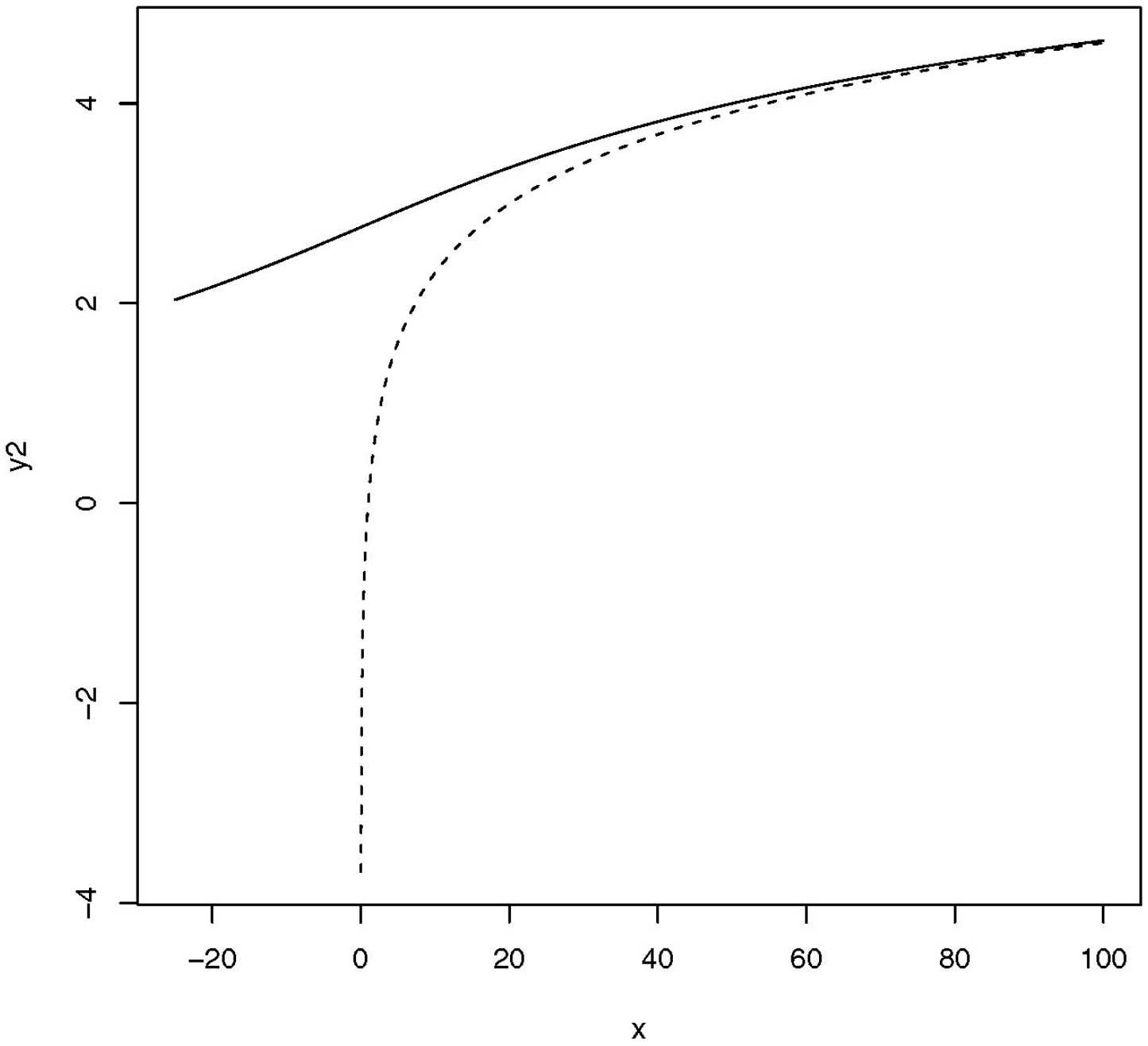
$$f(\mu) = \ln(\mu + \sqrt{\mu^2 + a^2/b^2}).$$

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

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

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


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

Alternative Transformations

- Similar results can be obtained with the started log transformation $\log(y + c)$ or the log linear hybrid transformation.

- The constants need to be carefully chosen to maintain approximate constant variance.

- Log transforms on non-background-corrected data amounts to the same thing as a started log for the background-corrected data with a particular choice of the started-log constant. This may or may not work well depending on the exact details of the quantification in the assay itself.

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. The measured expression is derived from the mean glog-transformed PM probes.

Steps in the Analysis

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

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

- Summarize each probe set using perfect match probes.

- Normalize the chips additively. In some cases, intensity-based normalization may be needed.

- The transformation should remove systematic dependence of the gene-specific variance on the mean expression, 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.

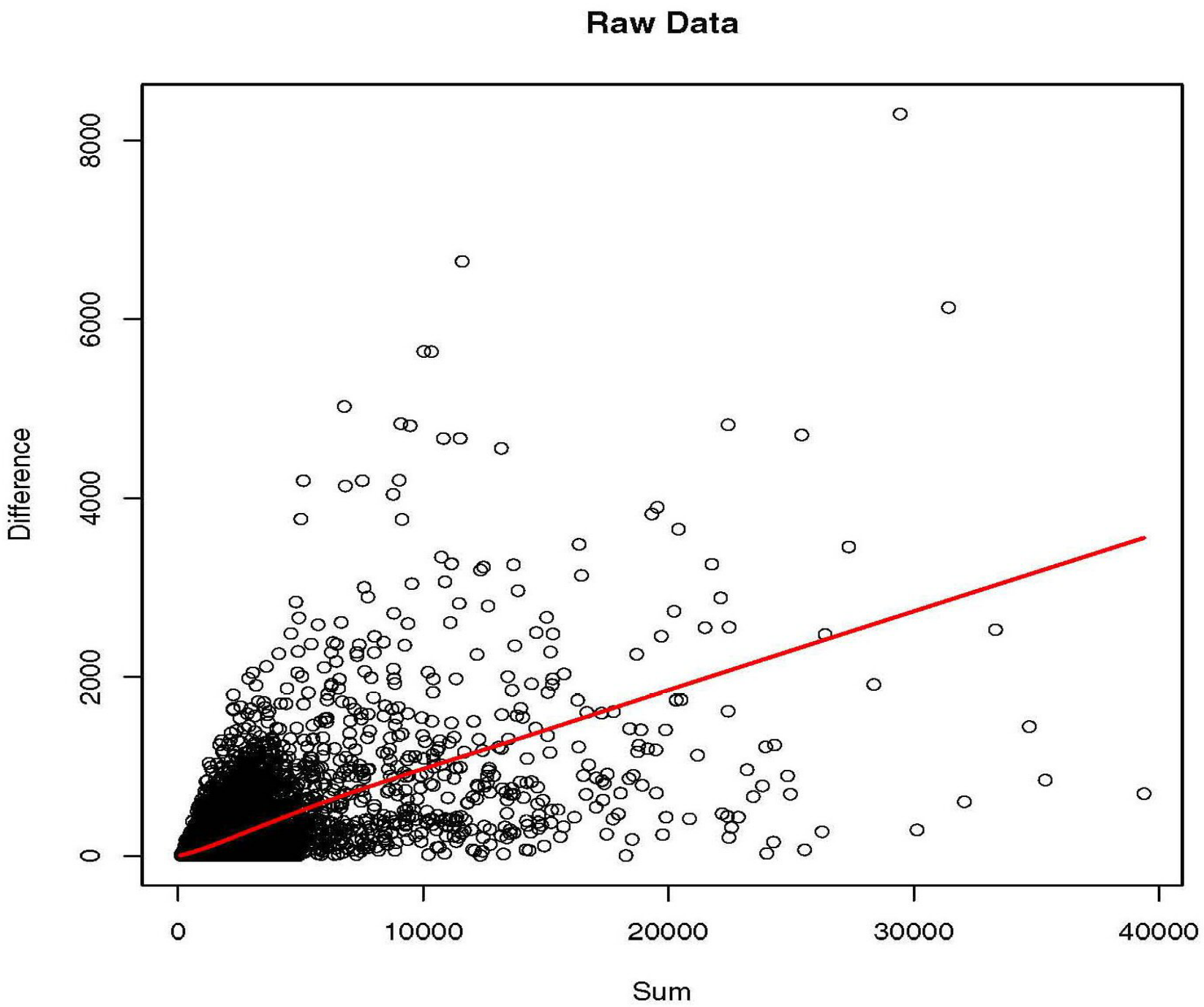
- Adjust p-values for multiplicity using, for example, the False Discovery Rate method.

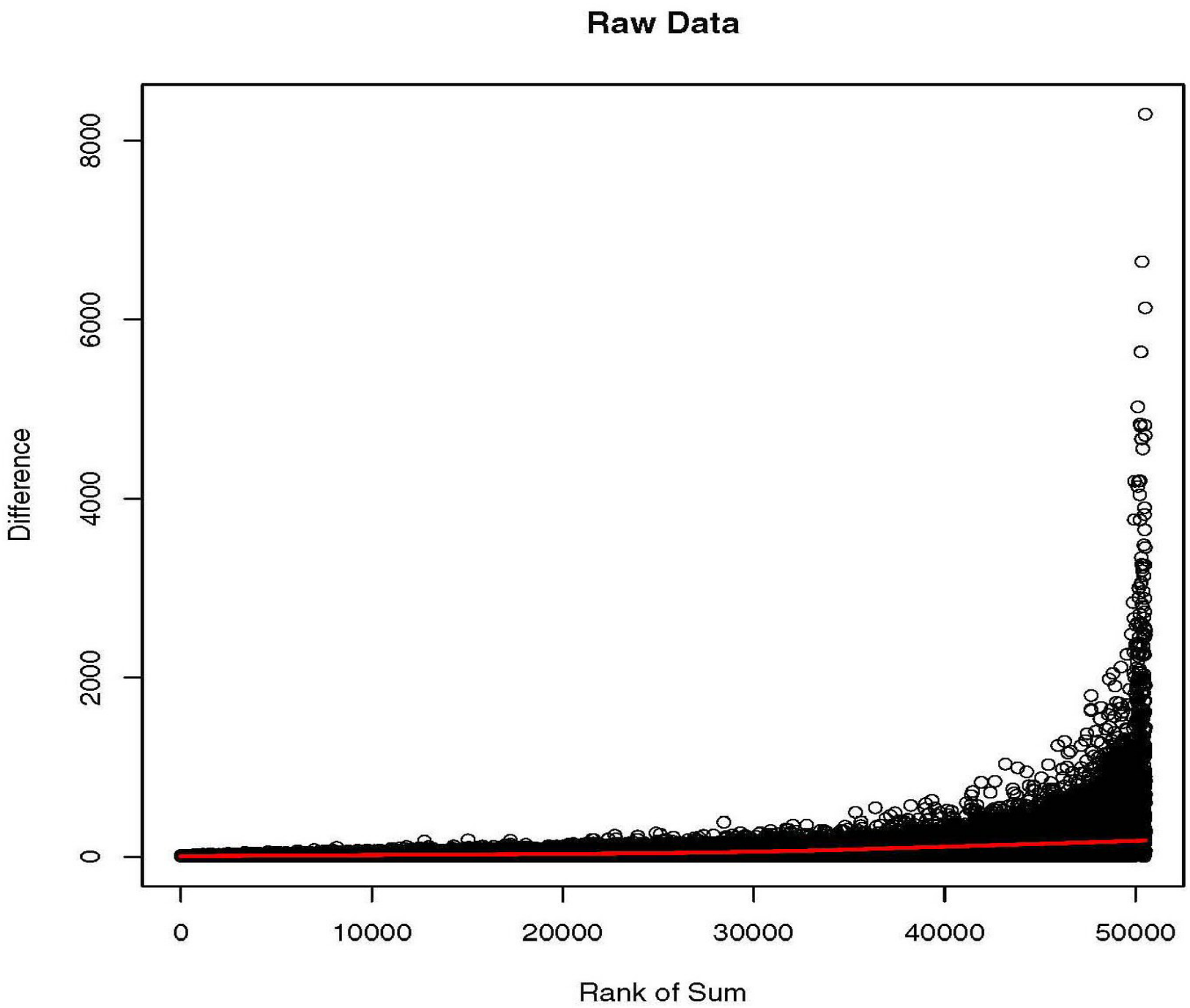
- Provide list of differentially expressed genes

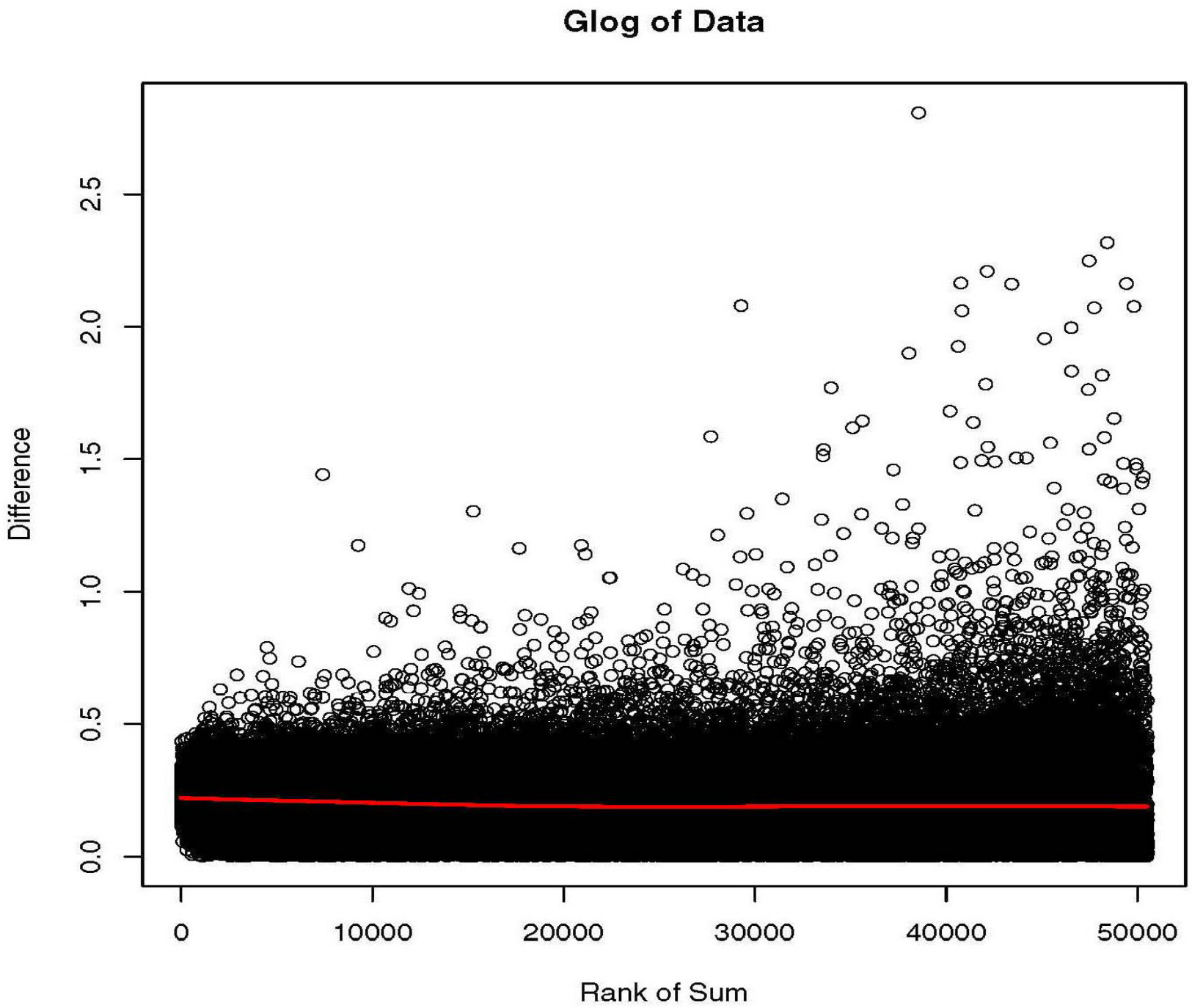
- Investigate identified genes statistically and by biological follow-up experiments.

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
Sum
0
2000
4000
6000
8000
0
10000
20000
30000
40000


Raw Data
Difference
0
2000
4000
6000
8000
0
10000
20000
30000
40000
50000
Rank of Sum


Glog of Data
Difference
0.0
0.5
1.0
1.5
2.0
2.5
0
10000
20000
30000
40000
50000
Rank of Sum

The model we use is (Kerr, Martin, and Churchill)

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

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

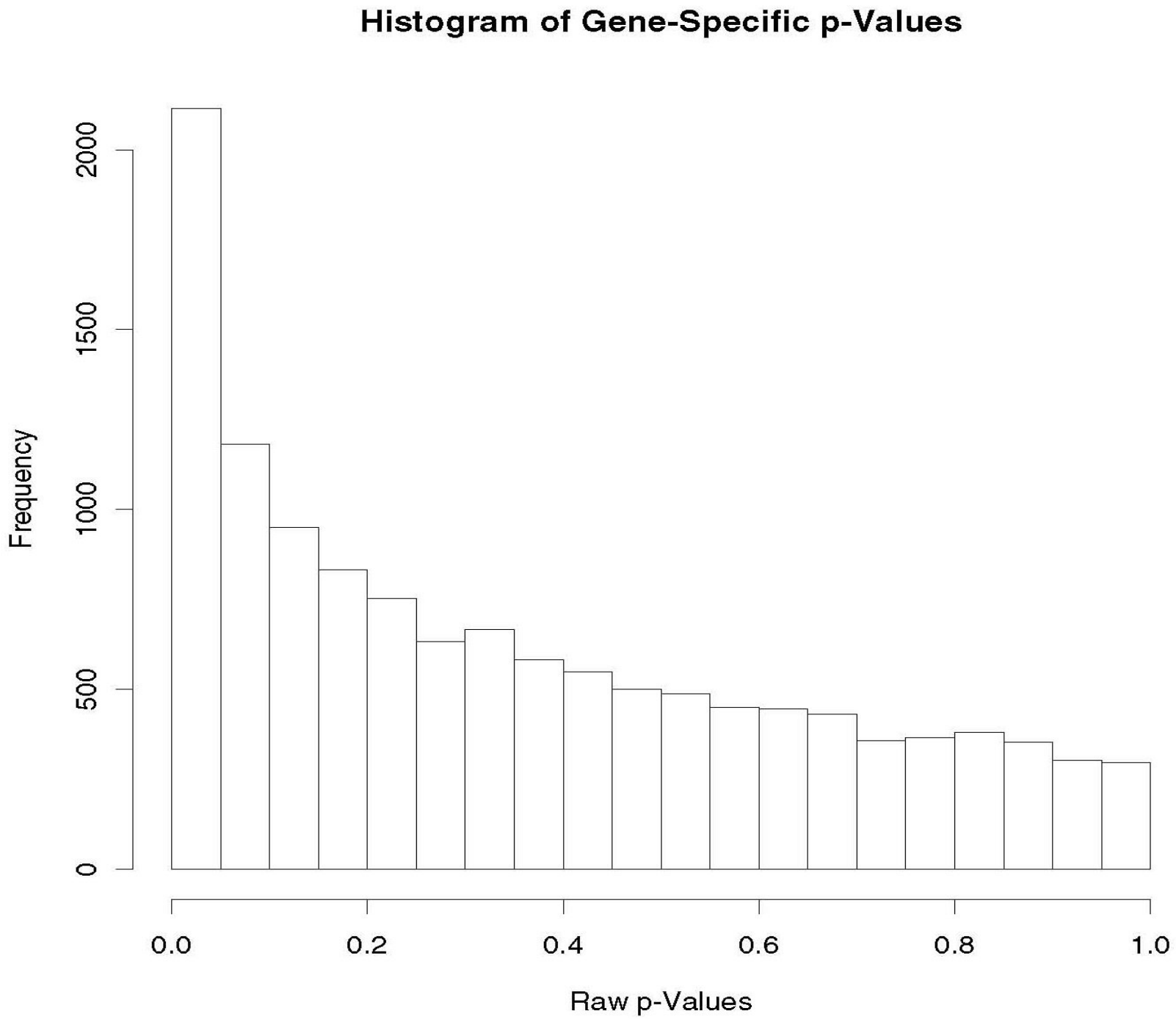
- The maximum likelihood parameter values for the data y will be the same as the minimum MSE parameter values from the model for the transformed Jacobian-corrected data z.

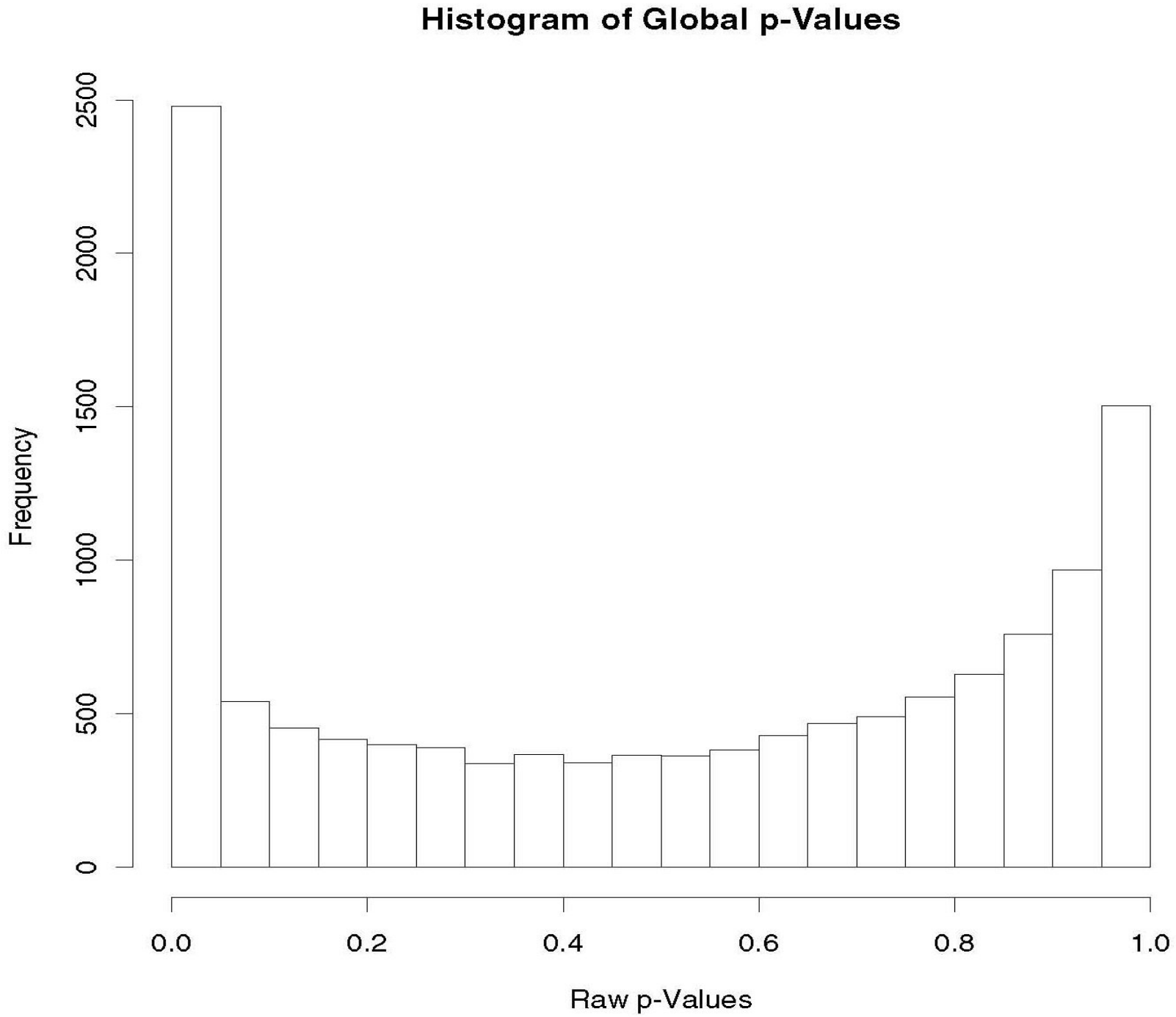
We can't fit the model for the full data set using most linear model software. The X matrix has $12{,}625 \times 8 = 101{,}000$ rows and $12{,}625 \times 4 + 7 = 50{,}507$ columns. The $X^\top X$ matrix is then 50,507 by 50,507, containing 2.55×10^9 8-byte reals, or 19GB!

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

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

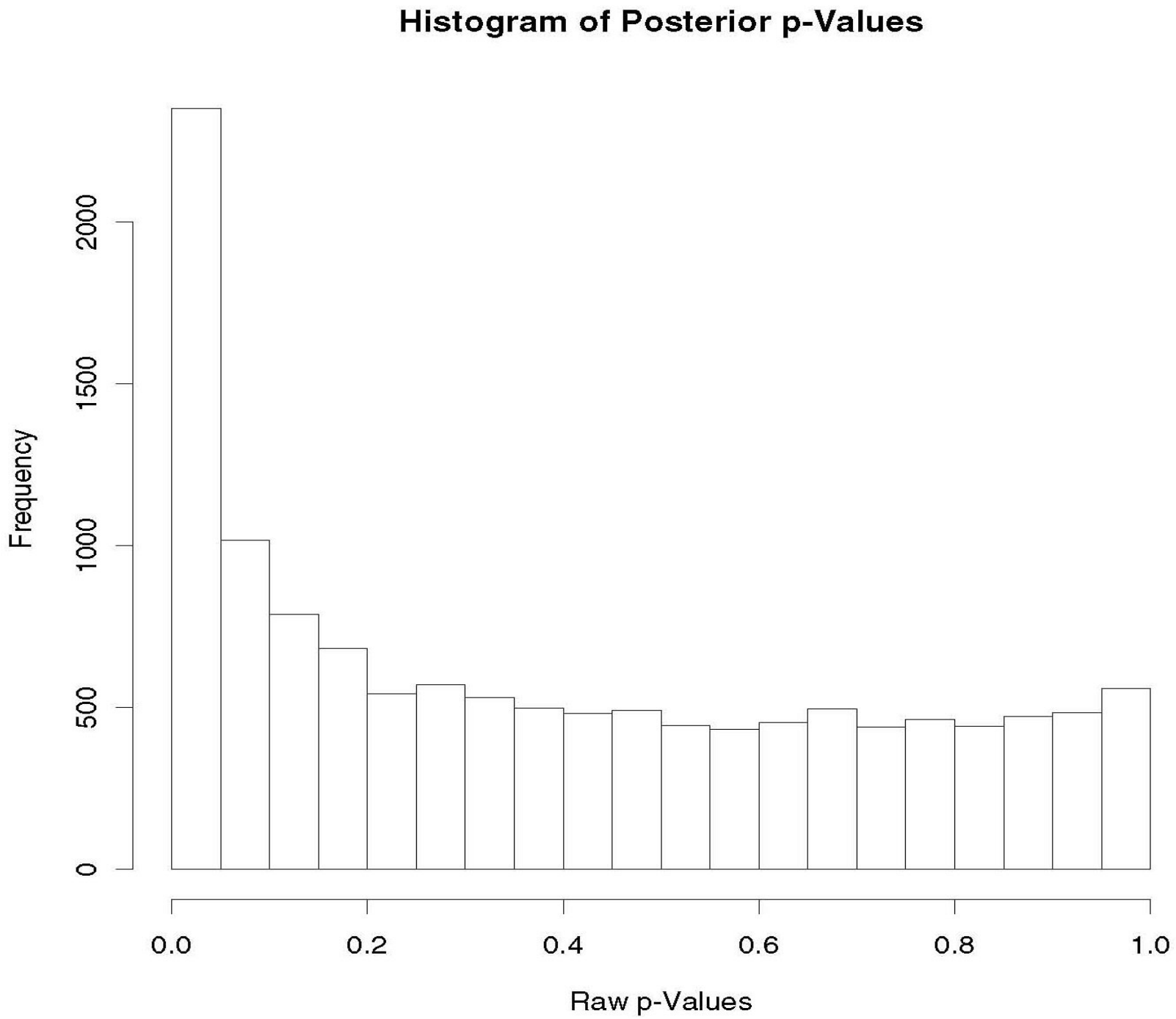
- We can use the overall 50,493df MSE as a denominator. This is powerful, but risky.


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 (see also Wright and Simon 2003).

- The overall MSE is 0.01212. The variance across genes of the 4df estimates under homogeneity should be approximately $2\sigma^4/4 = (.01212)^2/2 = 7.35 \times 10^{-5}$ Instead, the variance is 80.37×10^{-5},which is more than 10 times larger.


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

MSE Source	TWER	FWER	FDR
Gene-Specific	2054	0	0
Global	4215	1029	2693
Posterior	2804	48	866

Metabolomics by NMR Spectroscopy

- Proton NMR spectroscopy produces a spectrum in which peaks correspond to parts of molecules.

- This can be used for single compounds to determine the structure.

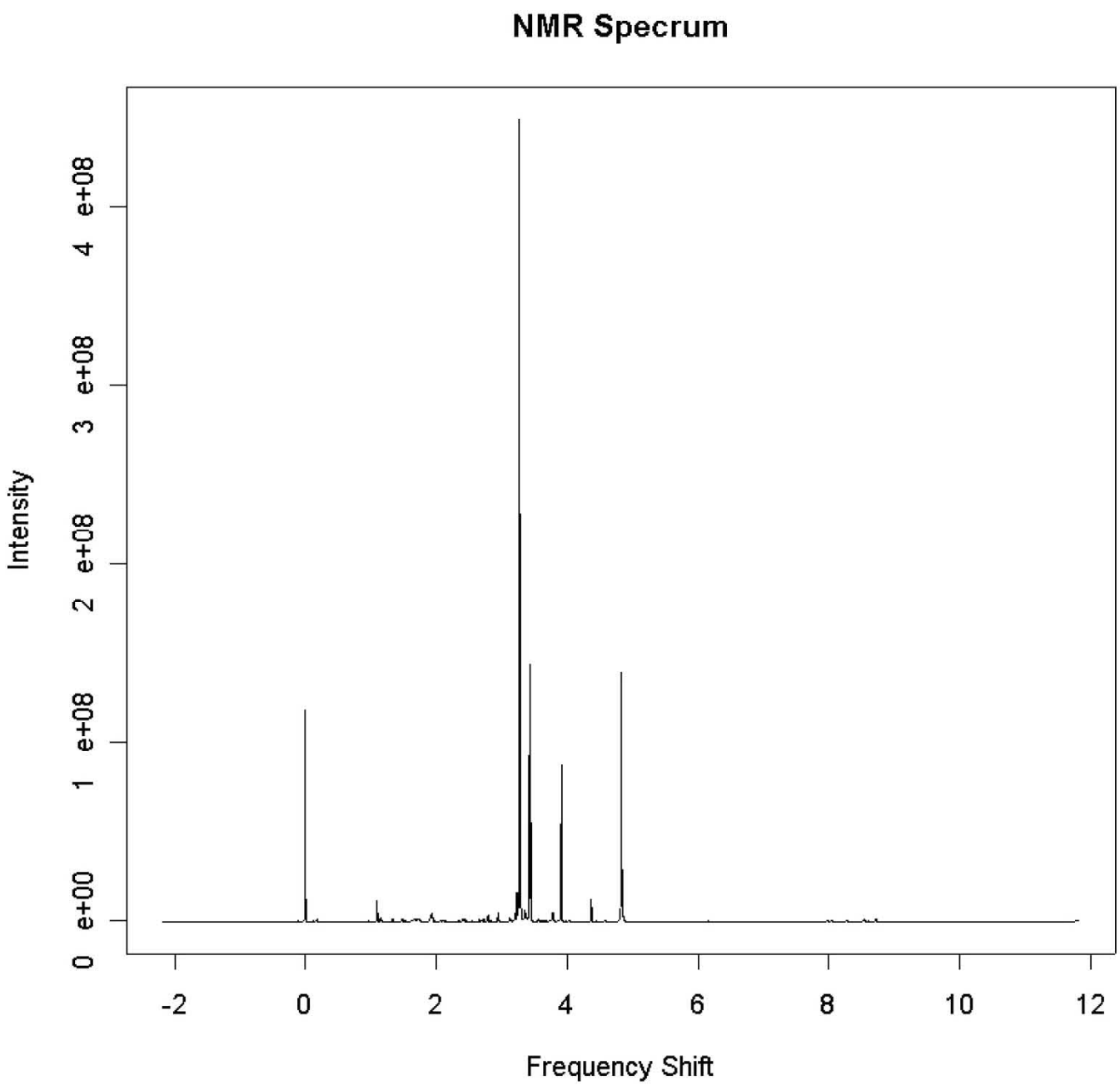
- Compounds often have specific signatures, so this can be used for compound identification (particularly by 2D NMR).

- For metabolomics work, one can use patterns in the spectra for discrimination/classification, and to identify regions of the spectrum which carry the discrimination information.

- Spectra need to be baseline corrected and peaks need to be aligned

- 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.


NMR Specrum
Intensity
0 e+00
1 e+08
2 e+08
3 e+08
4 e+08
-2
0
2
4
6
8
10
12
Frequency Shift

Variance Behavior of NMR Spectra

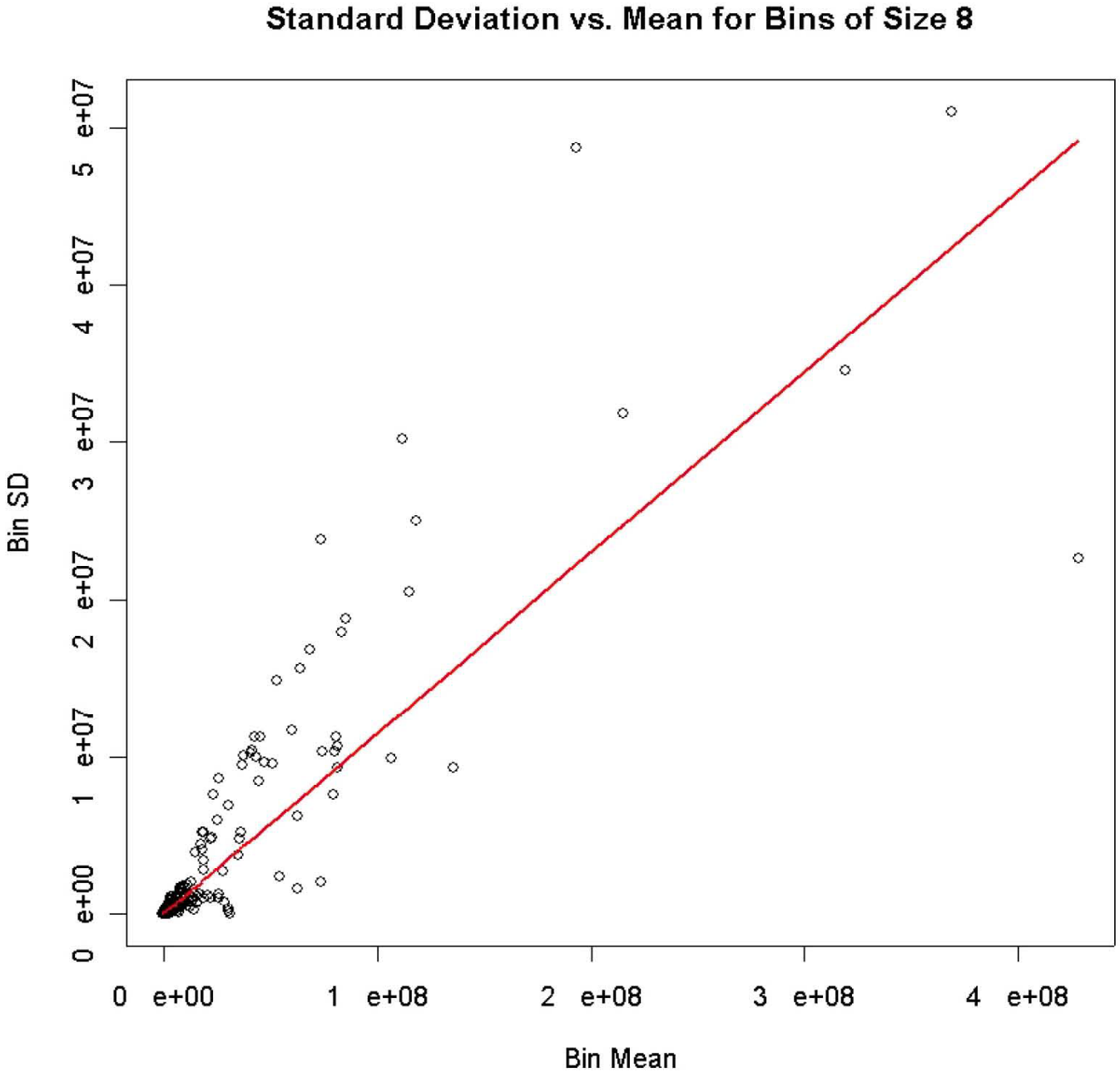
- We show an example spectrum of 65,536 points.

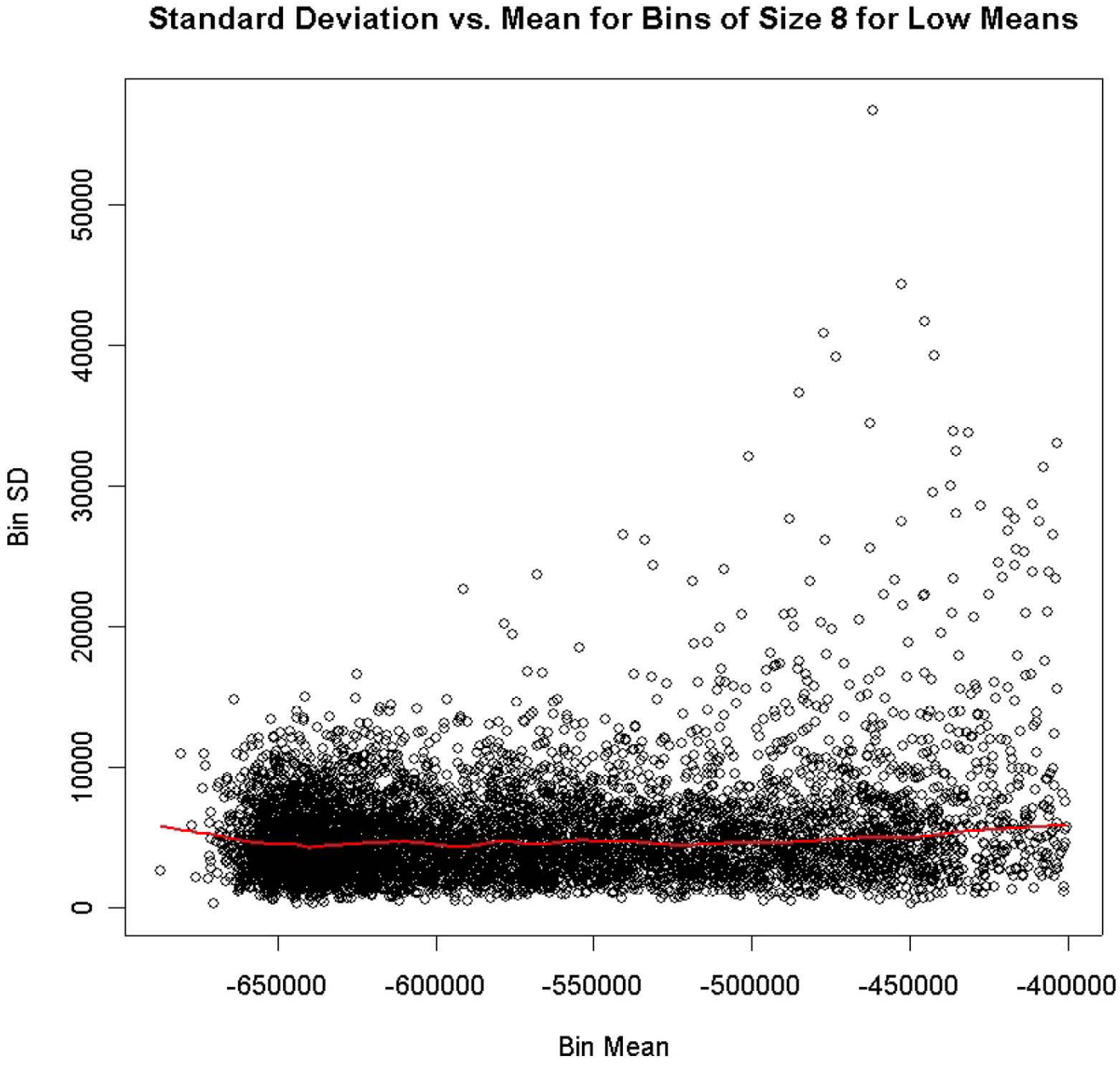
- We divide this into 8,192 bins of 8 points each and compute the mean and standard deviation within each bin.

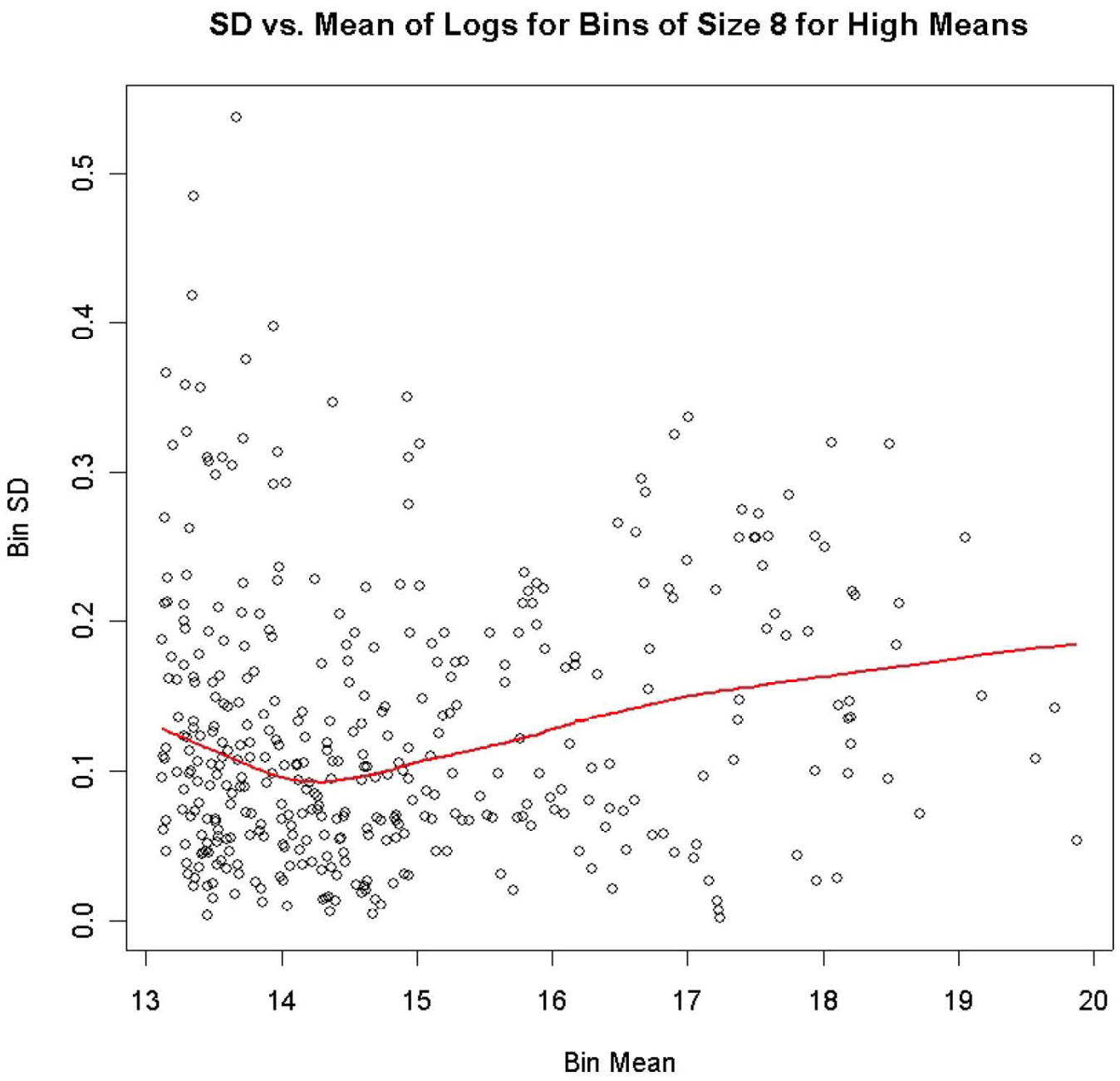
A model for the spectrum is

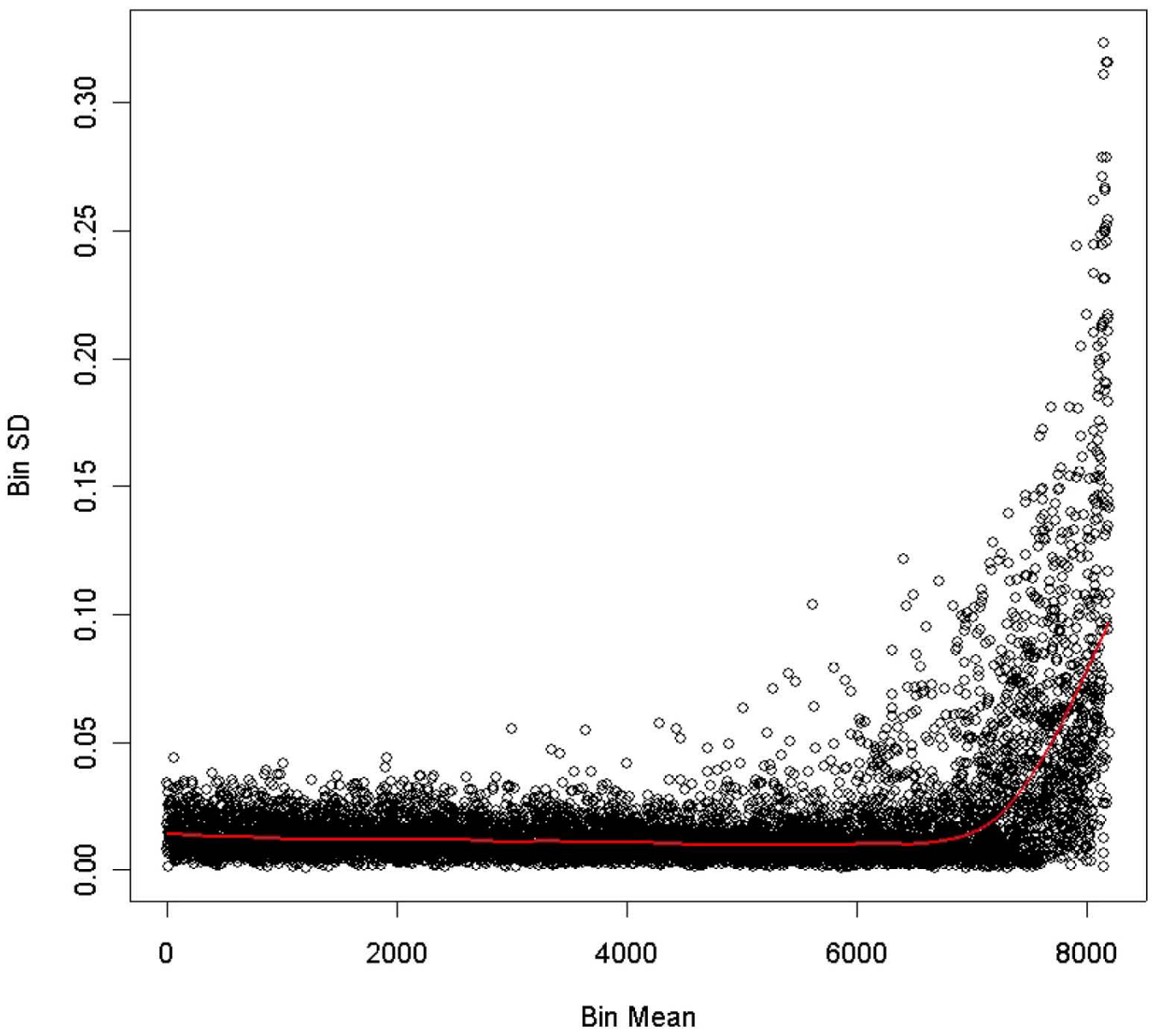
$$y_i = b_i + \mu_i e^{\eta} + \epsilon_i$$

Where b_i is the baseline, not presumed to be flat, μ_i is the true signal, and ϵ_i and η_i are measurement errors, not necessarily independent across nearby points.


Standard Deviation vs. Mean for Bins of Size 8
Bin SD
0 e+00
1 e+07
2 e+07
3 e+07
4 e+07
5 e+07
0 e+00
1 e+08
2 e+08
3 e+08
4 e+08
Bin Mean


Standard Deviation vs. Mean for Bins of Size 8 for Low Means
Bin SD
0
10000
20000
30000
40000
50000
-650000
-600000
-550000
-500000
-450000
-400000
Bin Mean


SD vs. Mean of Logs for Bins of Size 8 for High Means
Bin SD
0.0
0.1
0.2
0.3
0.4
0.5
13
14
15
16
17
18
19
20
Bin Mean


SD vs. Mean of Glogs for Bins of Size 8
Bin SD
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
2000
4000
6000
8000
Bin Mean

Baseline Estimation

- The baseline in NMR is arbitrary and needs to be removed before analysis, just as in mass spec.

- The baseline is less well behaved than for mass spec

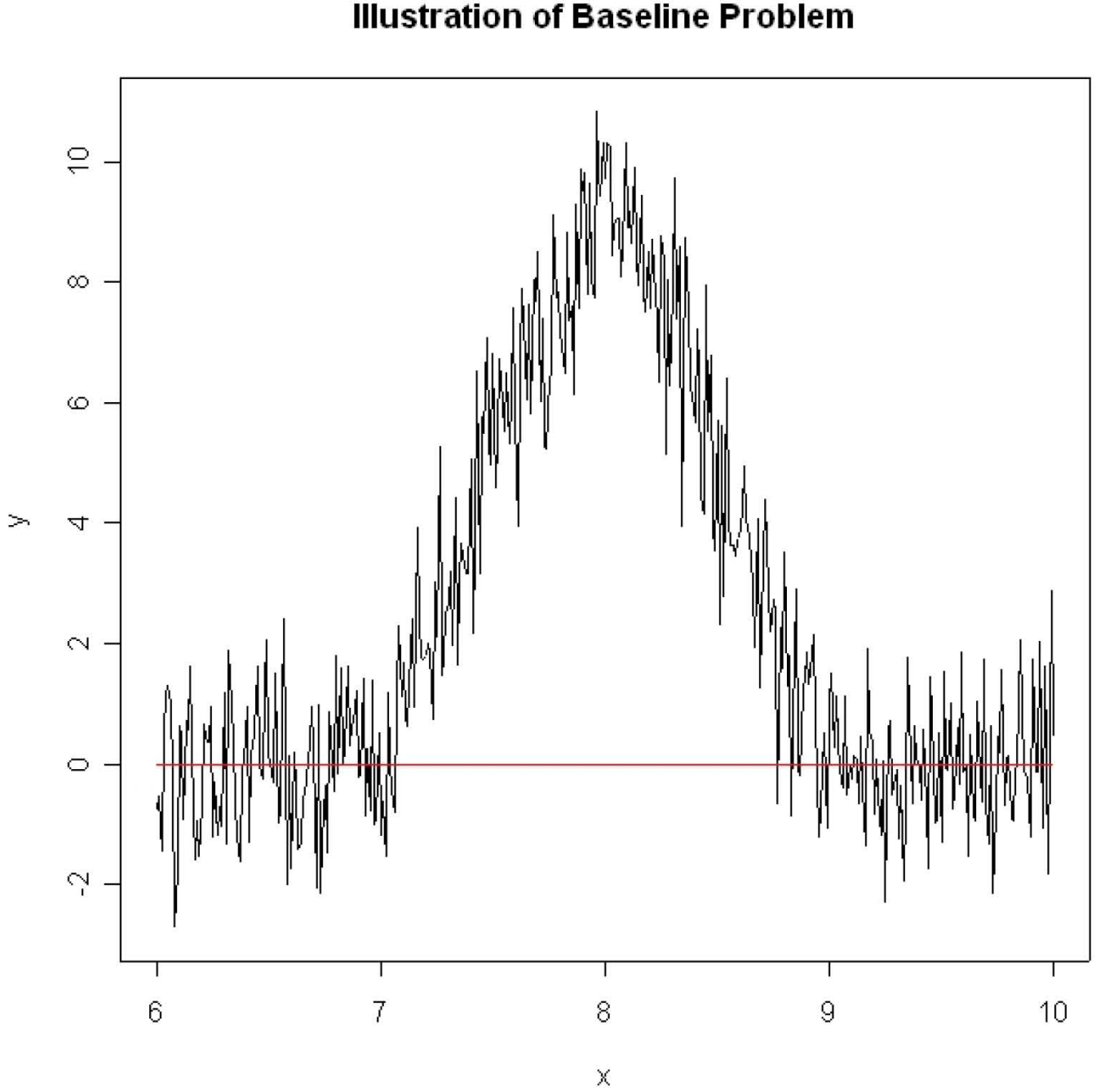
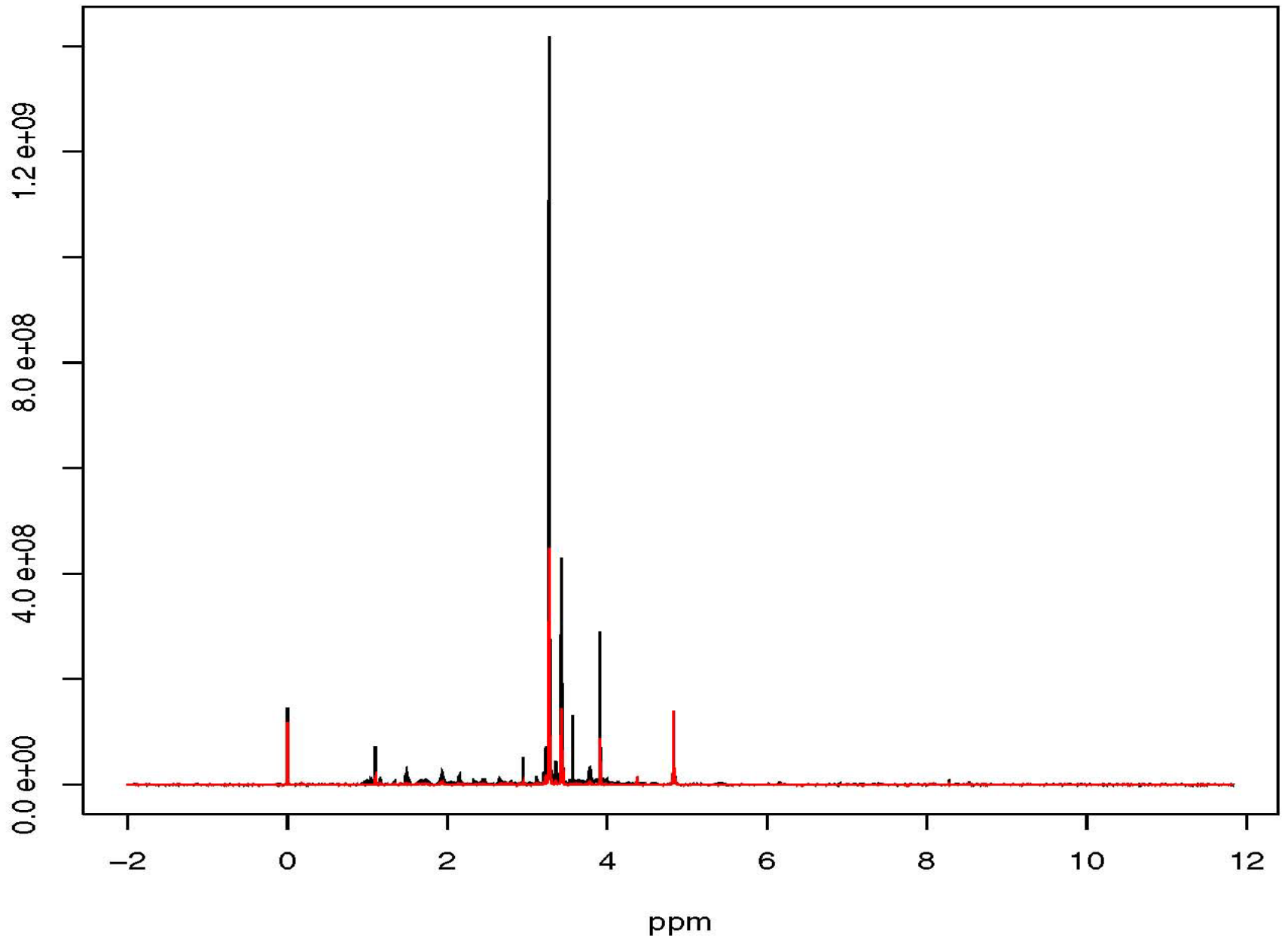


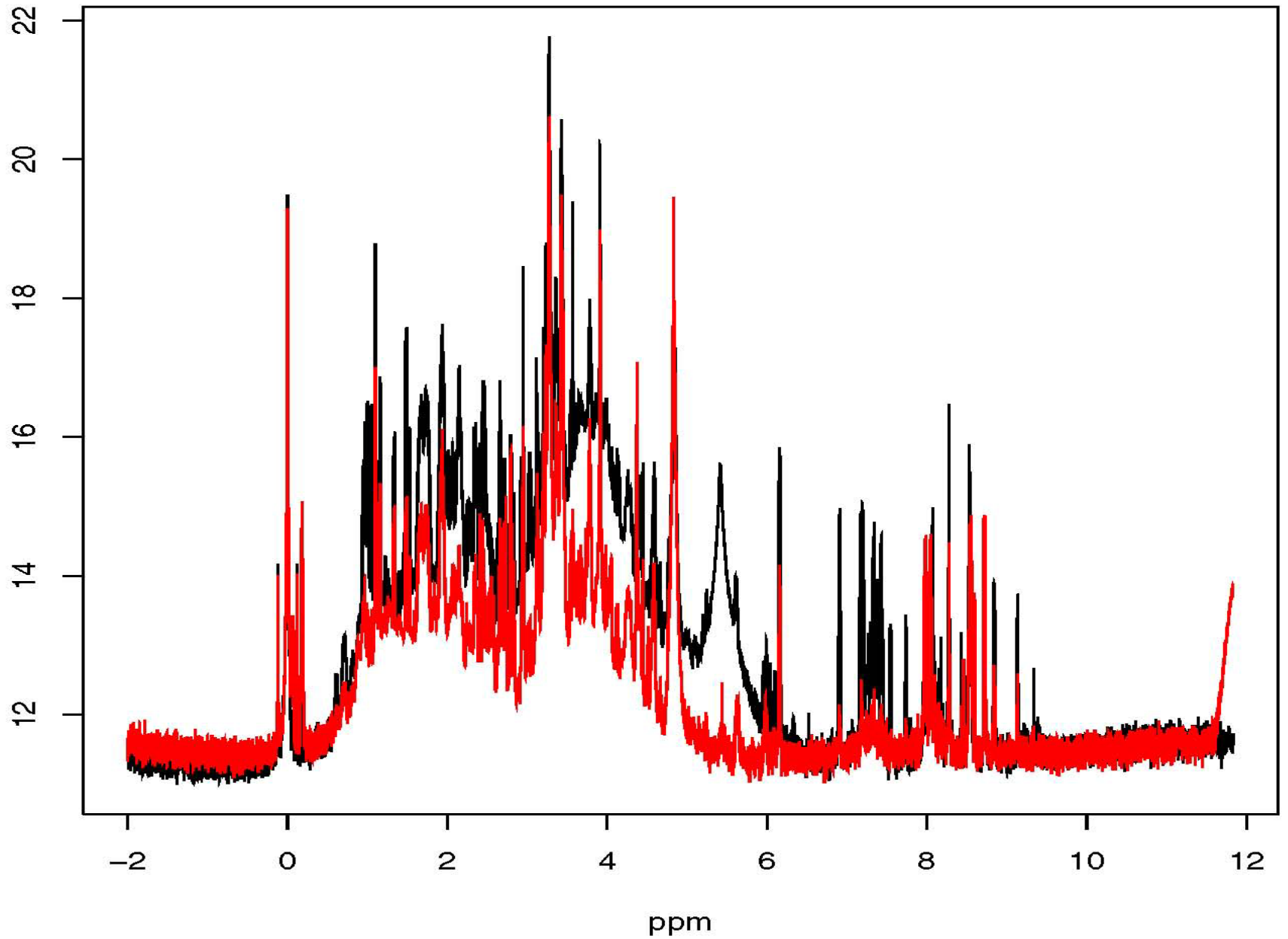
Illustration of Baseline Problem
y
x
-2
0
2
4
6
8
10
6
7
8
9
10

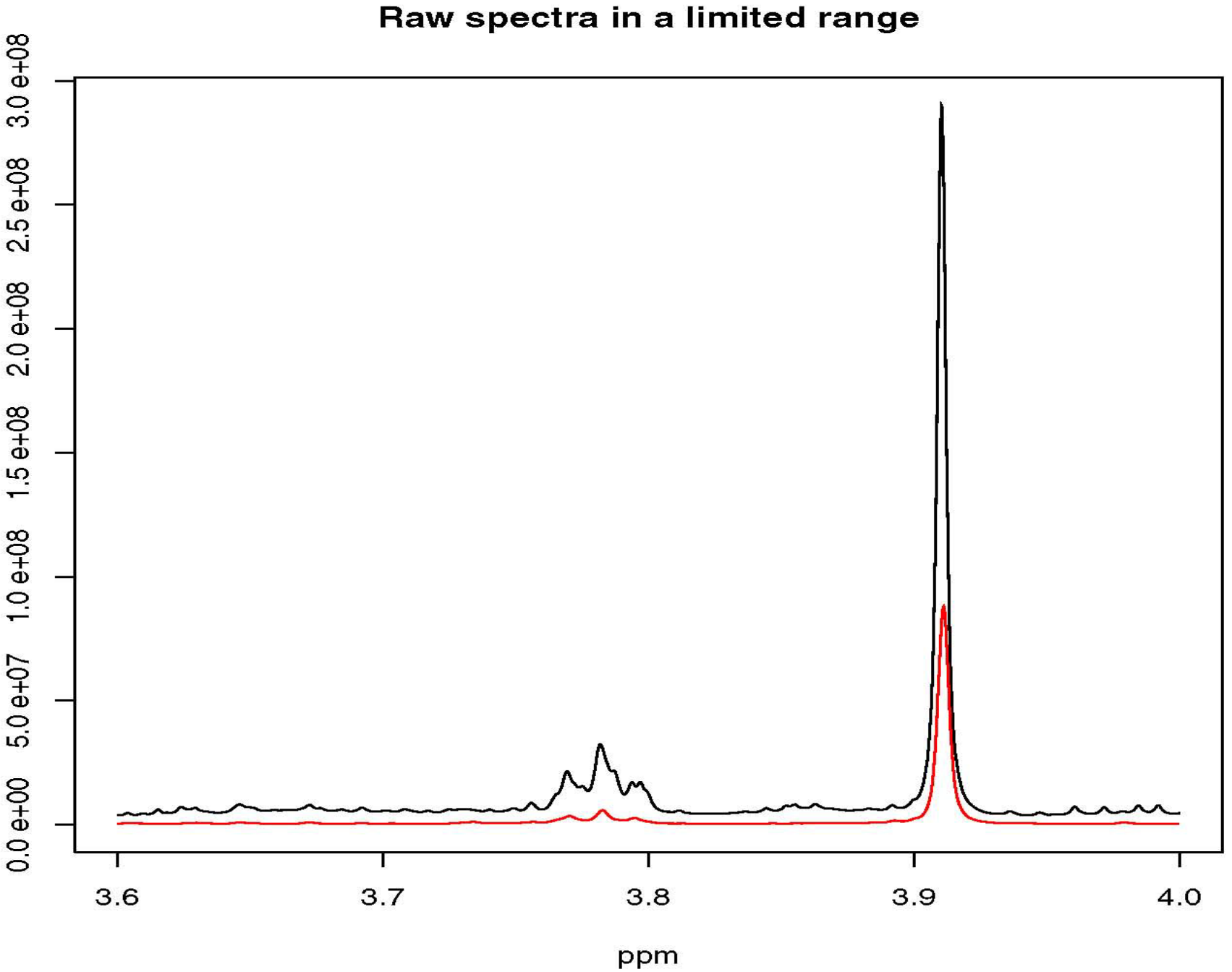
Our current best baseline estimation method solves the problem

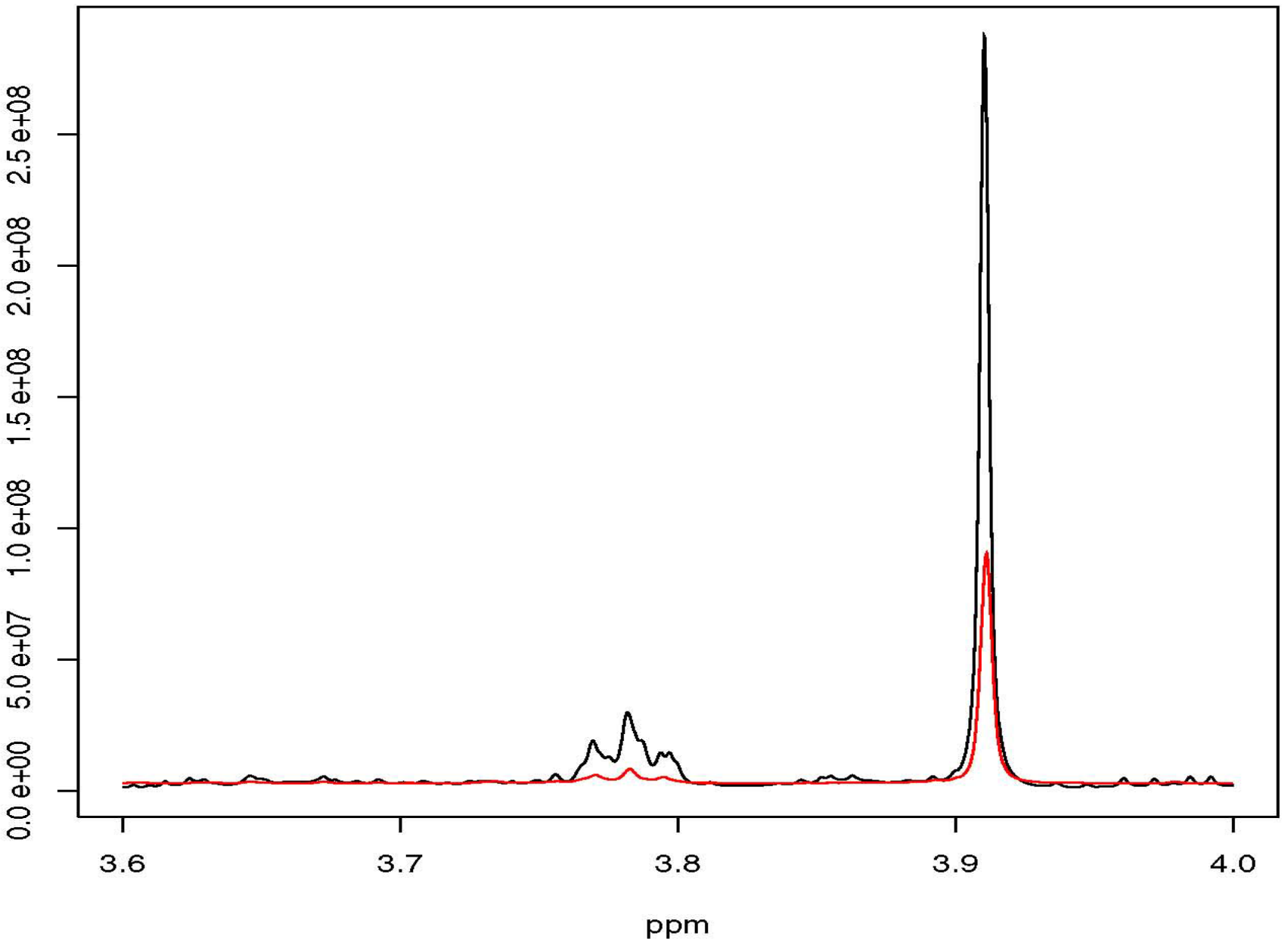
$$\max_{\{b_i\}} \left[\sum b_i - A \sum \left(\frac{y_i - b_i}{s_\epsilon} \right)^2 I(b_i - y_i) - B \sum (\Delta^2 b_i)^2 \right]$$

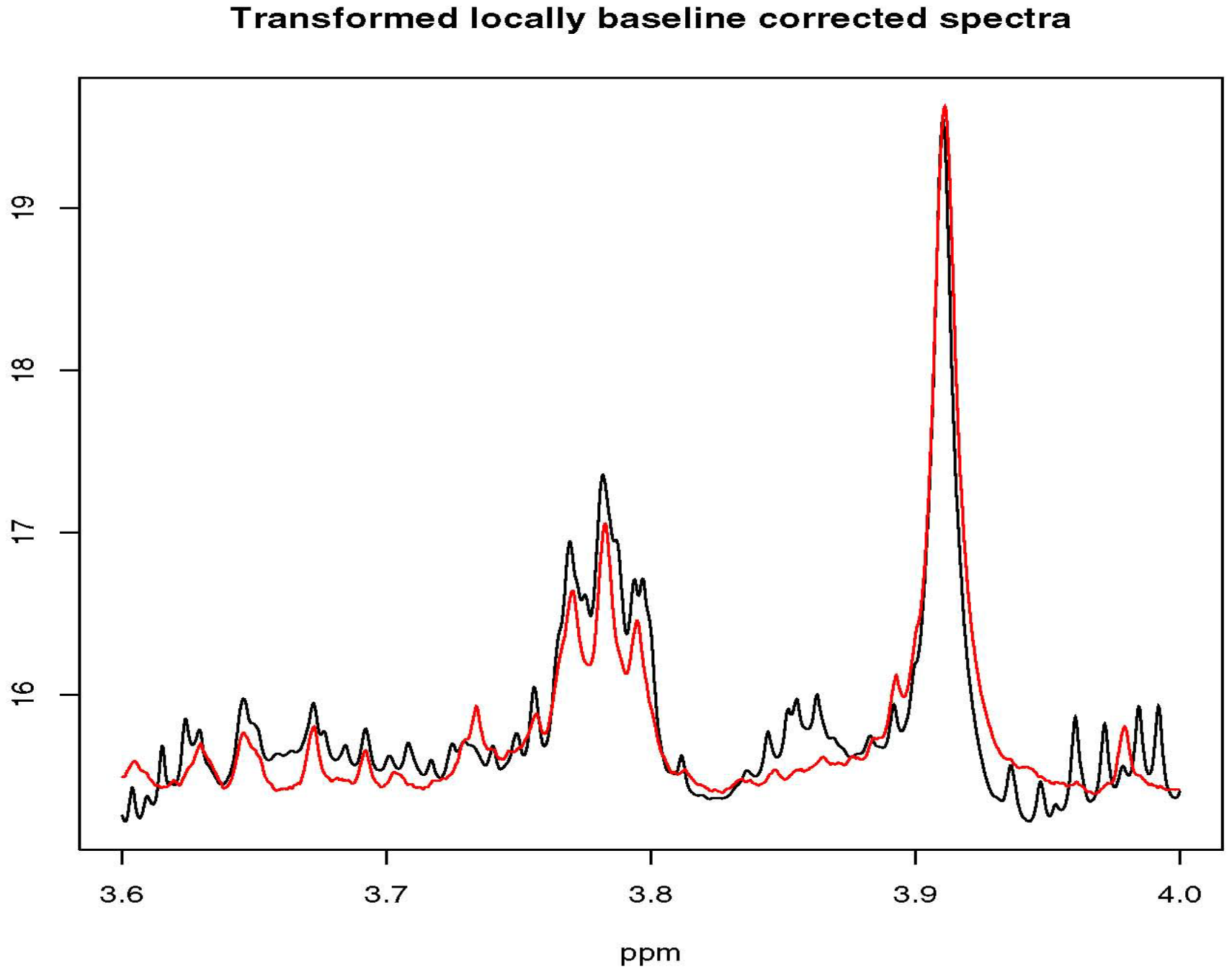
This is solved exactly by repeated solution of a banded linear system of $65,536 \times 65,536$, not using splines or other functional models. The constant B is chosen to achieve approximately unbiased estimation of the baseline if there is no signal.

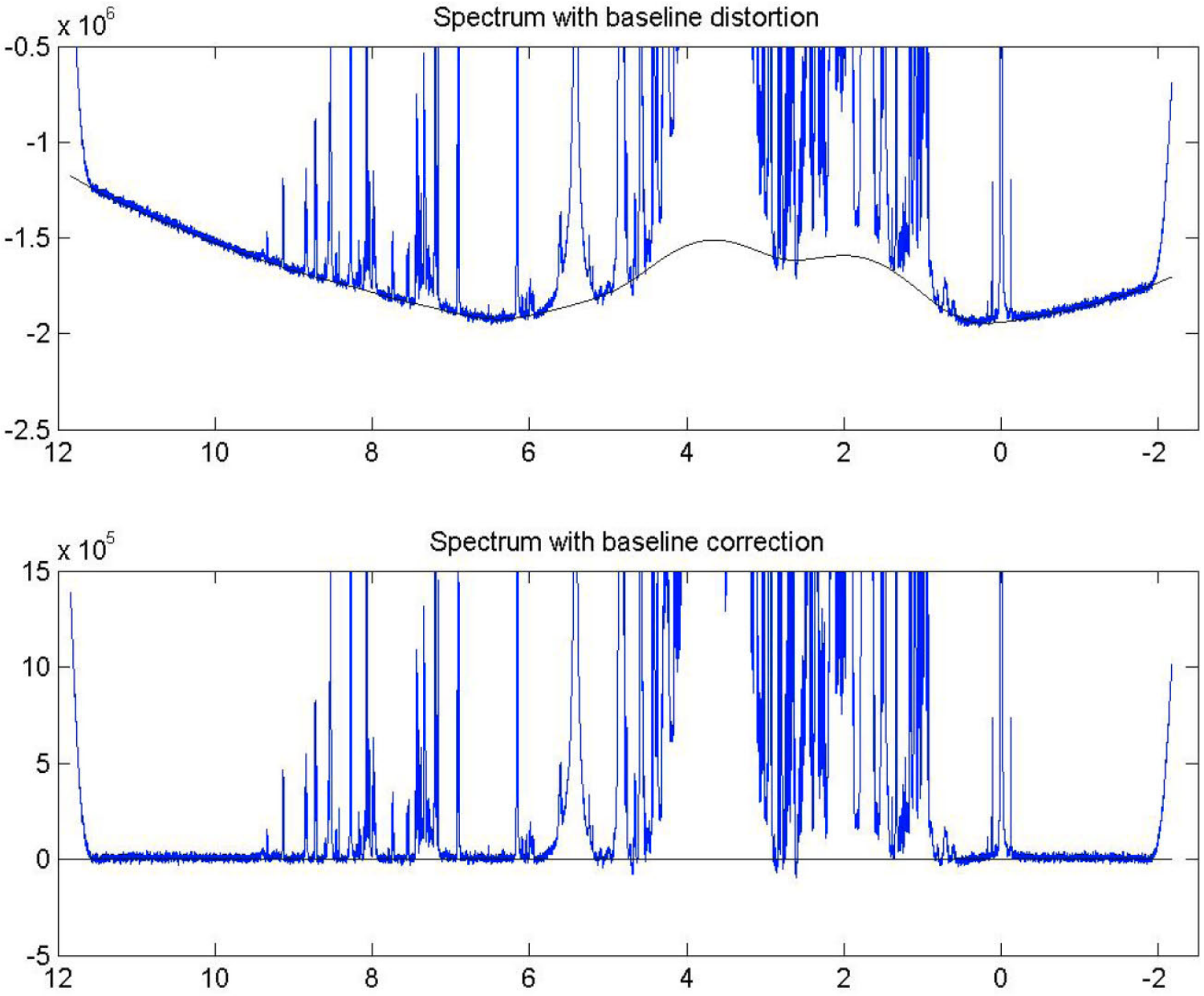

Raw baseline–corrected spectra
0.0 e+00
4.0 e+08
8.0 e+08
1.2 e+09
−2
0
2
4
6
8
10
12
ppm

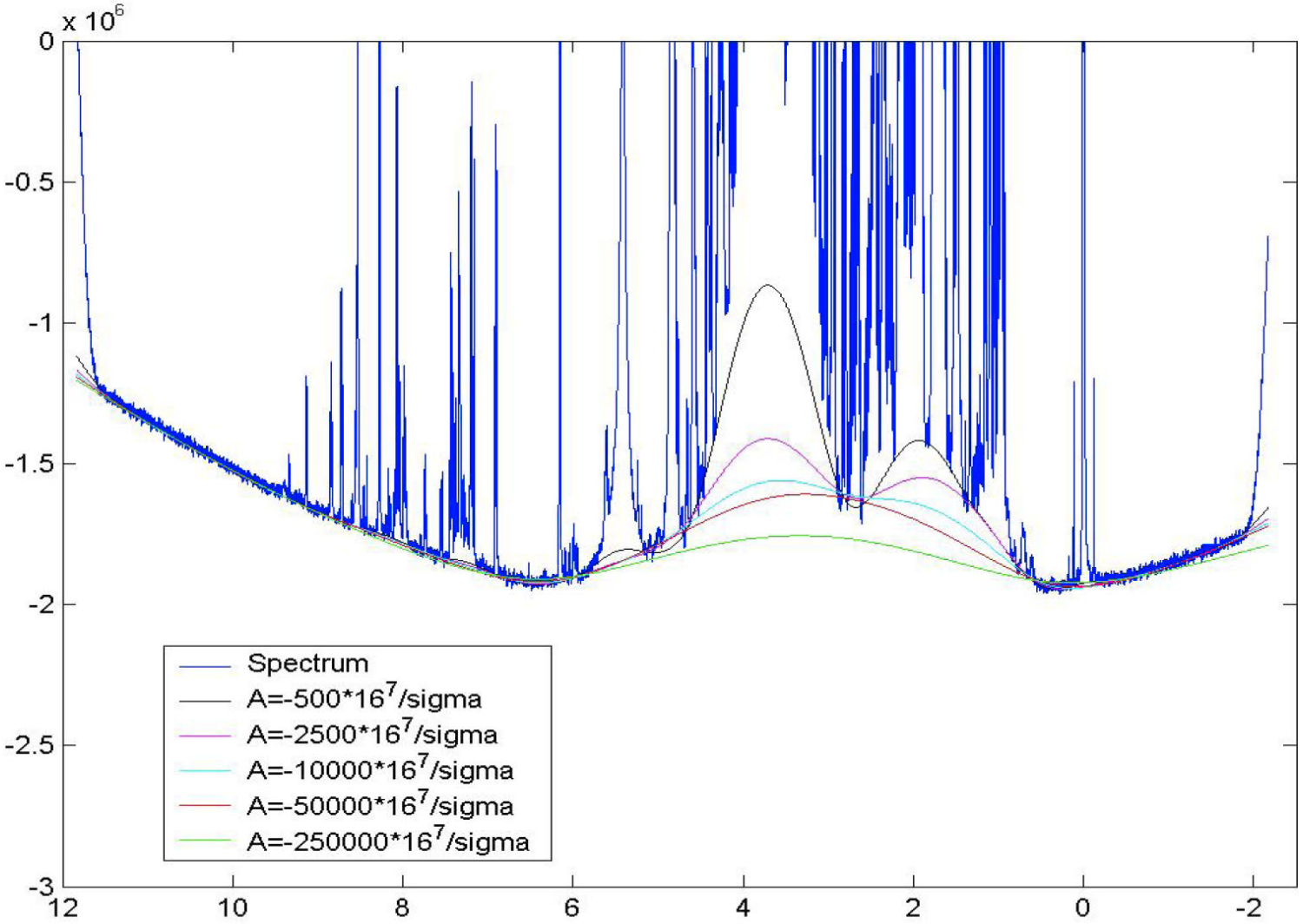

One glog transform of whole spectrum
22
20
18
16
14
12
−2
0
2
4
6
8
10
12
ppm


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


Spectrum with baseline distortion
x 10^6
-0.5
-1
-1.5
-2
-2.5
12
10
8
6
4
2
0
-2
Spectrum with baseline correction
x 10^5
15
10
5
0
-5
12
10
8
6
4
2
0
-2


x 10^6
0
-0.5
-1
-1.5
-2
-2.5
-3
12
10
8
6
4
2
0
-2
Spectrum
A=-500*16^7/sigma
A=-2500*16^7/sigma
A=-10000*16^7/sigma
A=-50000*16^7/sigma
A=-250000*16^7/sigma

Conclusion

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

- We have presented a model for variability 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 applying these methods to many types of data such as proteomics by 2D PAGE and MALDI-TOF, metabolomics by LC/MS and NMR spectroscopy, and GC lipid metabolomics. The variables measured are a large number of peak heights or areas, or a large number of binned spectroscopic values. And of course to gene expression data.

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

Acknowledgements

IDAV Faculty
Sue Geller (Texas A&M)
David Woodruff
Research Staff and Postdocs
Jian Dai
Lexin Li
Parul Purohit
John Tillinghast
Students and Former Students
Shagufta Aslam (UCD)
Blythe Durbin (UC Berkeley)
Wen-Ying Feng
Johanna Hardin (Pomona College)
Dan Li
Shuang Liu
Danh Nguyen (UC Davis)
Machelle Wilson (Univ. Georgia)
Yuanxin Xi
Jingjing Ye
Jufen Zhou
Lei Zhou
UC Davis Collaborators
Matt Bartosiewicz
Alan Buckpitt
Satya Dandekar
Jeff De Ropp
Bruce German
Dorothy Gietzen
Zelanna Goldberg
Jeff Gregg
Paul Hagerman
Bruce Hammock
Rivka Isseroff
Carlito Lebrilla
Kent Pinkerton
Bob Rice
Pam Ronald
Ralph deVere White
Outside Collaborators
Doug Hawkins (University of Minnesota)
Wolfgang Huber (German Cancer Research Center, Heidelberg)
Larry Kauvar (Trellis Bioscience)
Robert Nadon (McGill University)
Dan Solomon (Scripps Institute)
Mark Viant (University of Birmingham)
Martin Vingron (MPI/Molecular Genetics, Berlin)
Steve Watkins (Lipomics)
Funding
NSF
NIEHS
NIH
US EPA
UC Davis MIND Institute