An Analysis of Normalization Methods

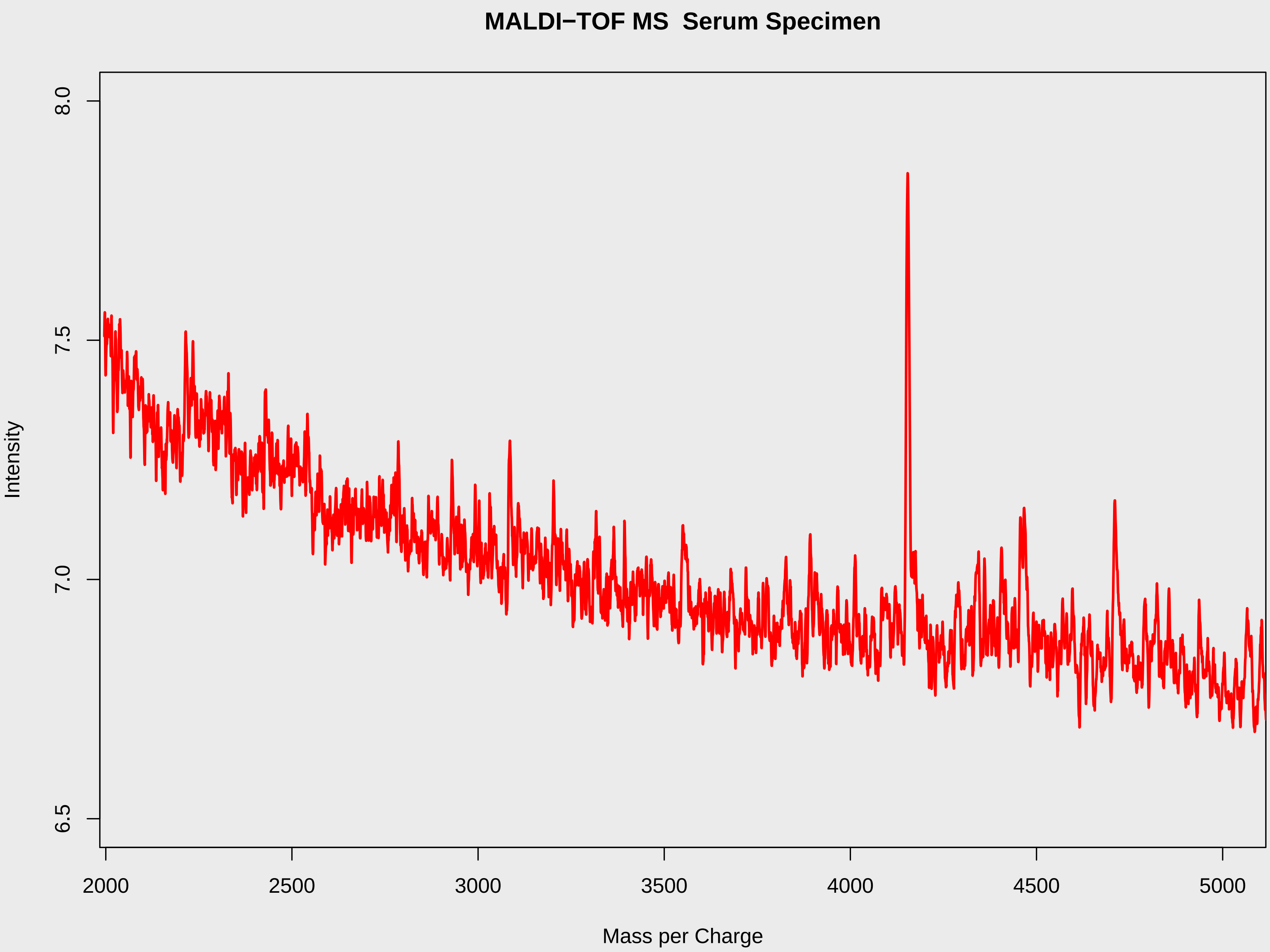
Bonnie LaFleur and Dean Billheimer

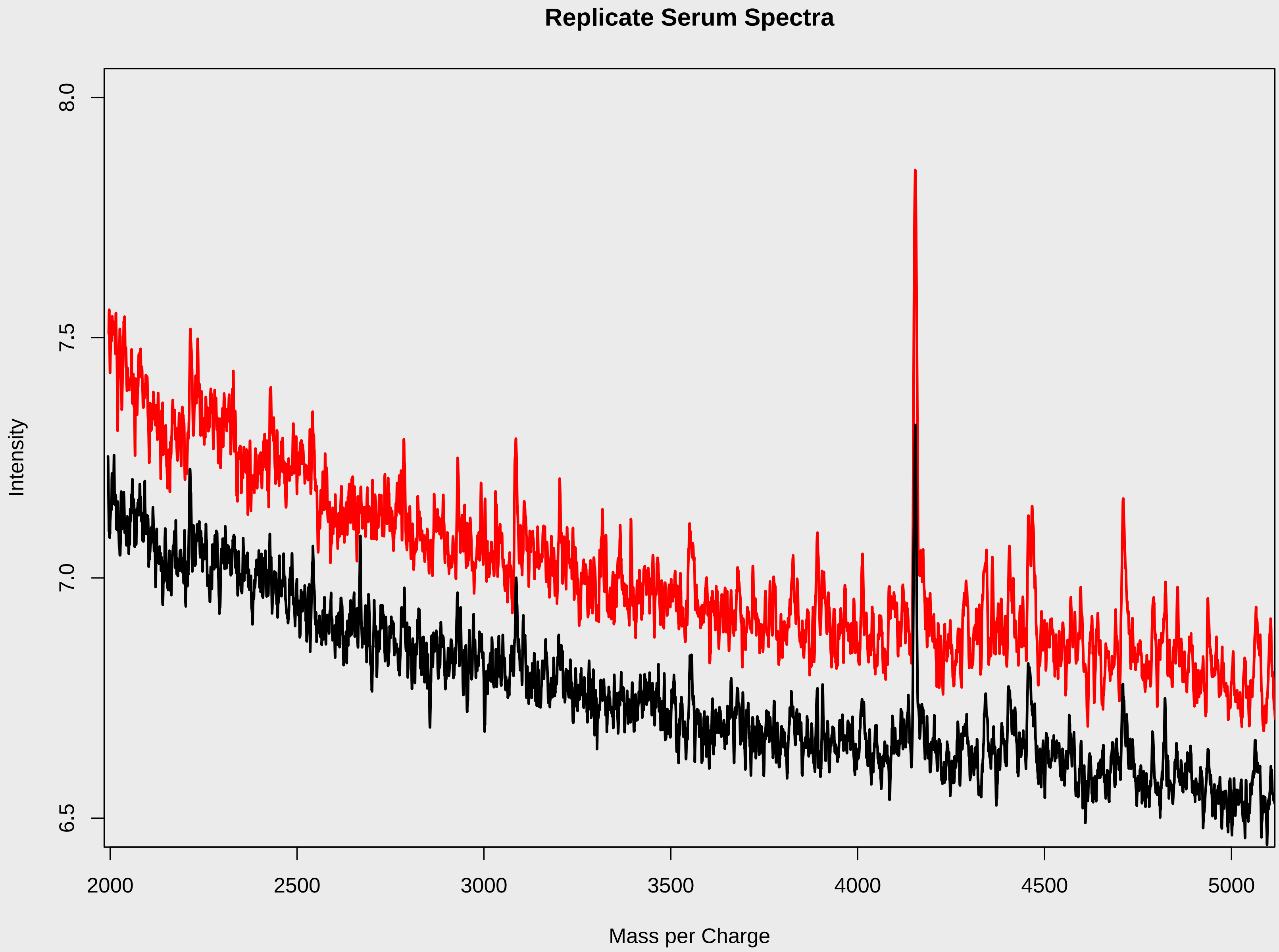
Departments of Pediatrics

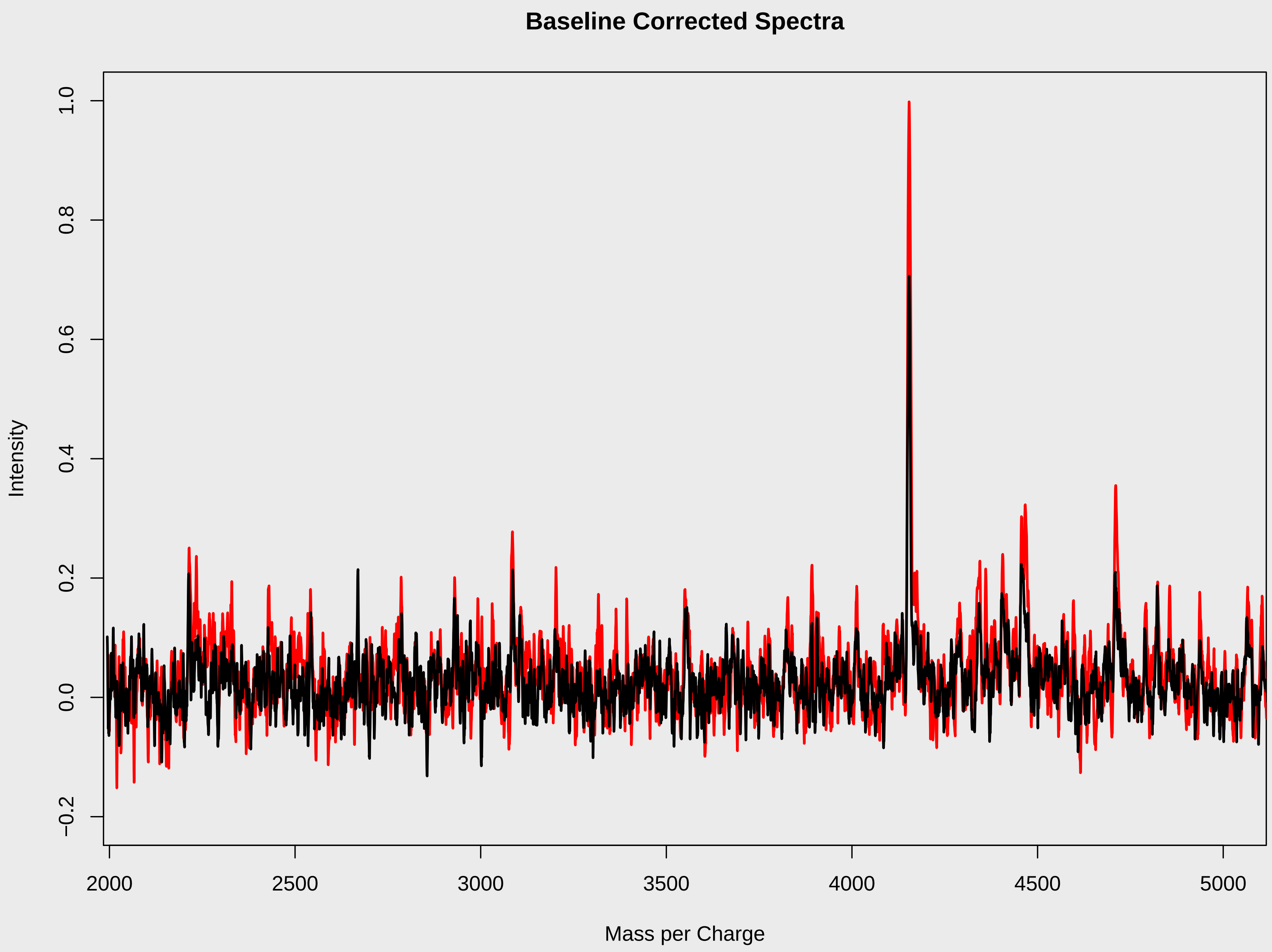
and

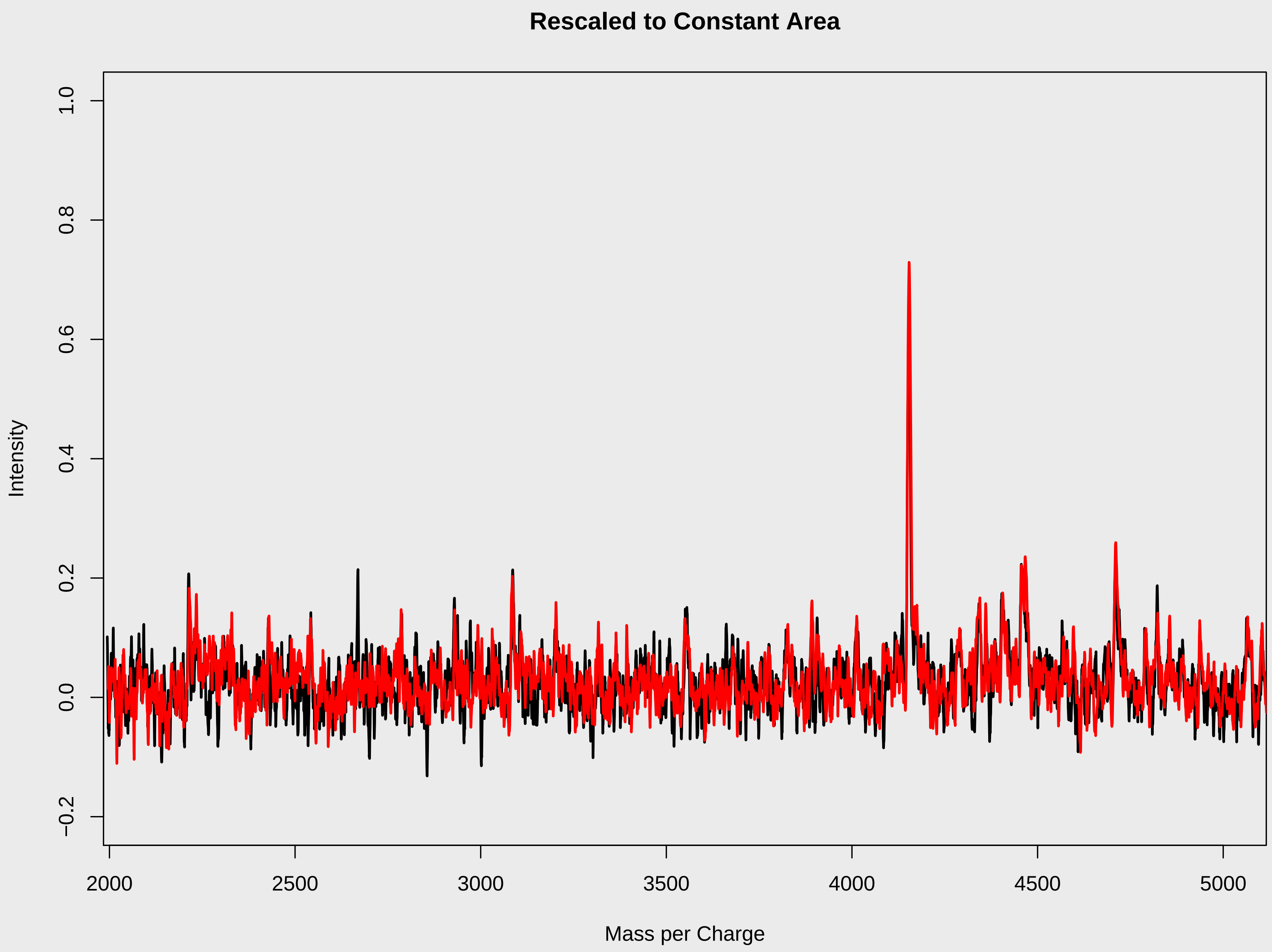
The Huntsman Cancer Center

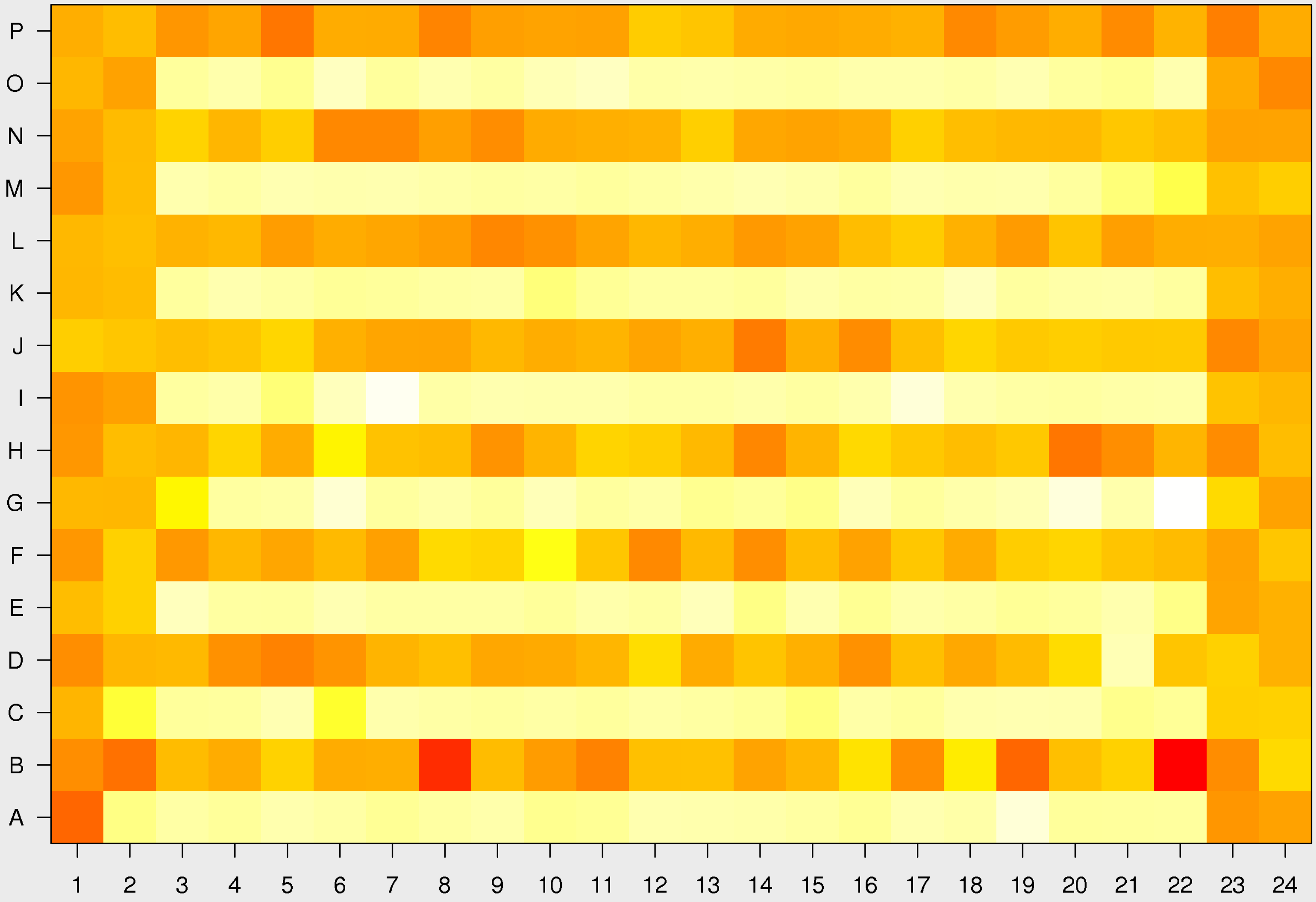
University of Utah


MALDI−TOF MS Serum Specimen
Intensity
8.0
7.5
7.0
6.5
2000
2500
3000
3500
4000
4500
5000
Mass per Charge


Replicate Serum Spectra
Intensity
8.0
7.5
7.0
6.5
2000
2500
3000
3500
4000
4500
5000
Mass per Charge

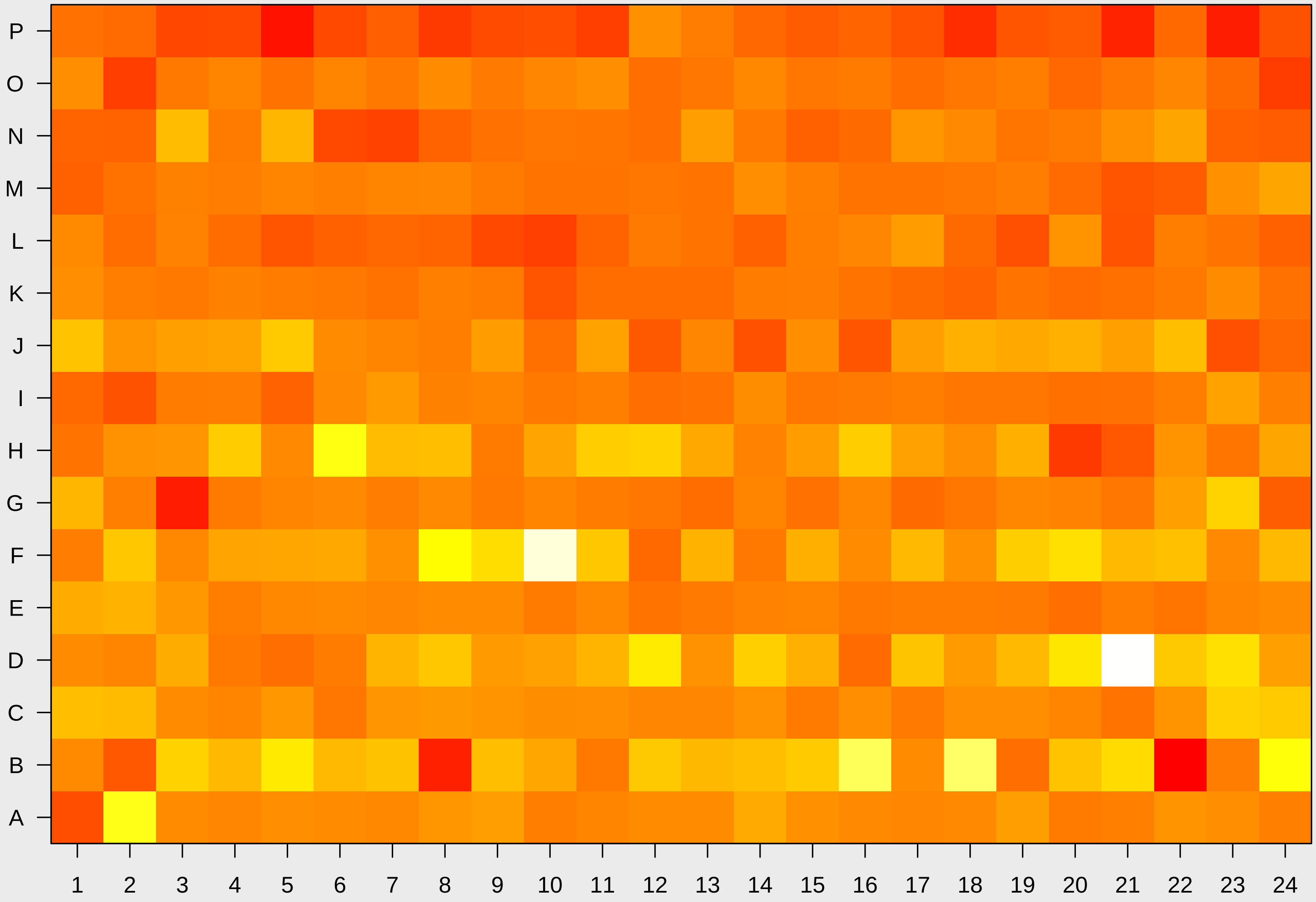

Baseline Corrected Spectra
Intensity
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
2000
2500
3000
3500
4000
4500
5000
Mass per Charge


Rescaled to Constant Area
Intensity
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
2000
2500
3000
3500
4000
4500
5000
Mass per Charge



P
O
N
M
L
K
J
I
H
G
F
E
D
C
B
A
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

Firefly_ Renilla_Prestwick 1 collection_1571_rep1 .ps



Data Preprocessing

- Smoothing or filtering
- Baseline or background correction
- Scaling (multiplicative)
- Nonlinear transformation

When the goal is to remove nuisance variation and (somehow) make the observations more “comparable”, we term this **normalization**.

Areas of Application

Normalization is used in a wide variety of measurement methods.

- DNA microarray
- Spectrometry (Raman, mass, others)
- Chromatography
- Quantitative gel electrophoresis
- lots and lots of others

It seems to be required when

- Highly multivariate measurements
- Indirect measurement (arbitrary units)
- Analog–Digital conversion (computer control)

Issues/Concerns for Scientists and Statisticians

- Details of normalization are not well described ("hidden" under preprocessing)
- Many algorithmic choices are made – often without the user's knowledge or control
- **These initial data manipulations are the most important steps in the data analysis**.
- How does one evaluate normalization methods? Choose between different methods?
- It is amazing that there is no standard theory (or even guidelines) for choosing normalization methods!

Road Map of next 40 Minutes

- What is Normalization?
- Cast Normalization as a Statistics Problem
- Identify Normalization Methods with Data Constraints
- Likelihood Based Metric
- Examples
- Summary
- Further Considerations

Think of this as a statistical normalization tutorial.

Normalization

Our goal is to develop a statistical theory of normalization.

- Characterize the normalization problem in a mathematical setting.
- Explicitly recognize the presence of variability.
- Identify important technical issues, and
- Provide an interpretable framework for addressing problems

The problem that normalization tries to correct:

The “ideal” data are given by

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

where $\mathbf{x}_i$, $\boldsymbol{\theta}_i$ and $\boldsymbol{\epsilon}_i$ are dimension p.

Through the measurement process we observe a corrupted version

$$\mathbf{y}_i = g(\mathbf{x}_i; \boldsymbol{\alpha}_i)$$

where $g(\cdot; \boldsymbol{\alpha}_i)$ is some function depending on nuisance parameters $\boldsymbol{\alpha}_i$.

Nuisance variation may include

- Baseline / background variation,
- Intensity scaling,
- A mean-variance dependence,
- Non-Gaussian error distributions.

Case of Multiplicative Nuisance Variation

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{bmatrix} = \begin{bmatrix} \alpha_1 & & & \\ & \alpha_2 & & \\ & & \ddots & \\ & & & \alpha_n \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}' + \boldsymbol{\epsilon}_1' \\ \boldsymbol{\theta}' + \boldsymbol{\epsilon}_2' \\ \vdots \\ \boldsymbol{\theta}' + \boldsymbol{\epsilon}_n' \end{bmatrix}$$

More compactly,

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

$\mathbf{Y}$ is an $n \times p$ matrix of observations,

$\boldsymbol{\Theta}$ is an $n \times p$ matrix of $p \times 1$ parameter vectors,

$\boldsymbol{\alpha}$ is a diagonal matrix of n nuisance parameters, and

$\boldsymbol{\eta}$ is an $n \times p$ error matrix.

Heuristic Normalization

Find invariant features in the data, and “normalize” so that these are constant for all observations.

Examples

- normalize a specific “signal” to known value (spike-in control).
- constant sum (mean) constraint,
- set observed maximum to 100%,
- quantile matching methods,
- choose a “representative” observation and transform to it.

Statistical Issues

- “Nuisance” variation is present for each (multivariate) observation
- Neyman–Scott (1948) – incidental parameter problem
- Model identification – if $\boldsymbol{\alpha}$ is unknown in $g(\cdot;\boldsymbol{\alpha})$
- Define normalization as a transformation of observed data to remove nuisance variation, and *identify the model*.

Model Identification

Each of the heuristic normalization strategies is an *ad hoc* choice to achieve model identification.

These can be written as rank one constraints.

- normalization to a specified signal $y_{it} = c$ for all i with t fixed
- constant sum normalization ($\equiv$ to mean constraint); $\mathbf{1}^{'}\mathbf{y}_i = c$
- normalization to the observed maximum; $\max_t(y_{it}) = c$

Key Idea: Each constraint defines a different normalization method.

Need a way to choose?

Need a comparative metric to aid selection of normalization method.

Desiderata include:

- easy to compute and interpret
- applicable across scales (e.g. $\sqrt{\ \ }$ or $\log(\ \ )$)
- coincide with graphical evaluation (when available)
- normalized data should be compatible with standard analysis methods

An Analysis of Transformations

Box and Cox, 1964 JRSS B

$$\begin{aligned} y &= \frac{x^\lambda - 1}{\lambda}, \quad \lambda \neq 0 \\ &= \log(x)\ , \quad \lambda = 0 \end{aligned}$$

But here’s what Box and Cox say:
We assume that for some unknown λ the transformed observations satisfy the full normal theory assumptions. The probability density for the untransformed observations, and hence the likelihood *in relation to these original observations*, is obtained by multiplying the normal density by the Jacobian of the transformation.

Key Idea: Use the likelihood to evaluate the transformation.

Likelihood Based Evaluation

Assume that after transformation (with some α) the data may be approximated by multivariate normal distribution. Then we may evaluate normalizations by the likelihood of the original data.

Let

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

then

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

where $\mathbf{J}(\hat{\alpha}, \mathbf{y})$ is the Jacobian of the transformation

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

How To Use the Proposed Evaluation Method

1. Select $g(\mathbf{x}_i; \boldsymbol{\alpha}_i)$ that describes your problem. For the multiplicative model this is $\mathbf{y}_i = \alpha_i \mathbf{x}_i$.
2. Choose an identifying constraint.
3. The constraint *defines* the normalization method (defines $\hat{\alpha}_i$)
4. Compute the transformed data $\mathbf{z}_i$.
5. Compute the log-likelihood of the original data, $\mathbf{Y}$
6. Repeat 2–5 for your favorite normalizations.
7. Larger likelihoods are better (for constraints with the same rank).

Technical Difficulties

- Constraints used for model identification induce singularity into the variance-covariance matrix.
- Use a generalized inverse to evaluate the likelihood (see e.g. Mardia, Kent and Bibby, 1979).

Simple Example

Generate frauda as follows:

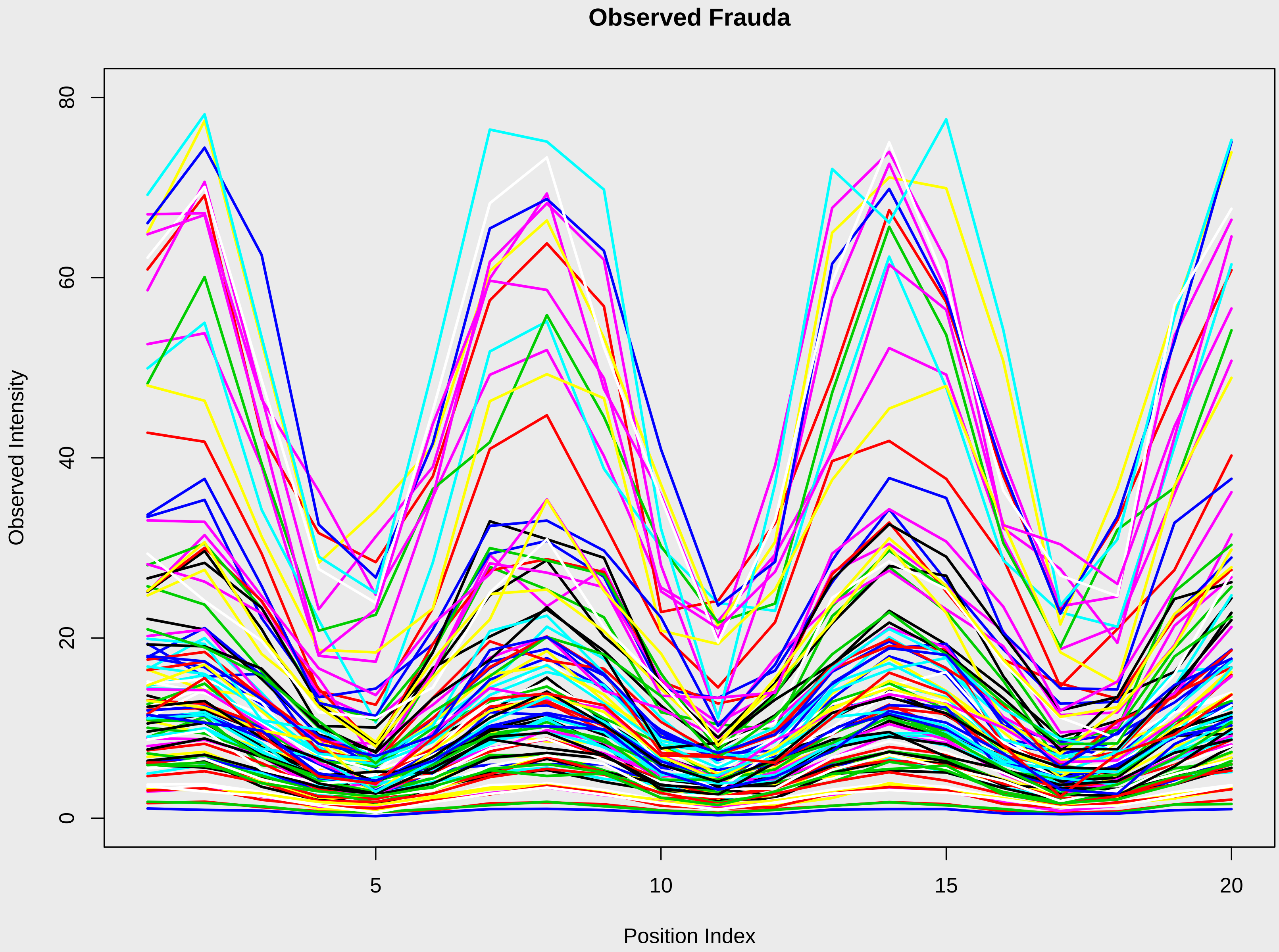
$$\boldsymbol{\mu}_t = 10 + \sin(t) \quad \text{where } t = 1, \ldots, 20.$$

$$\mathbf{y}_i = \alpha_i \left(\boldsymbol{\mu}_t + \boldsymbol{\epsilon}_i \right) \quad \text{where } i = 1, \ldots, 100.$$

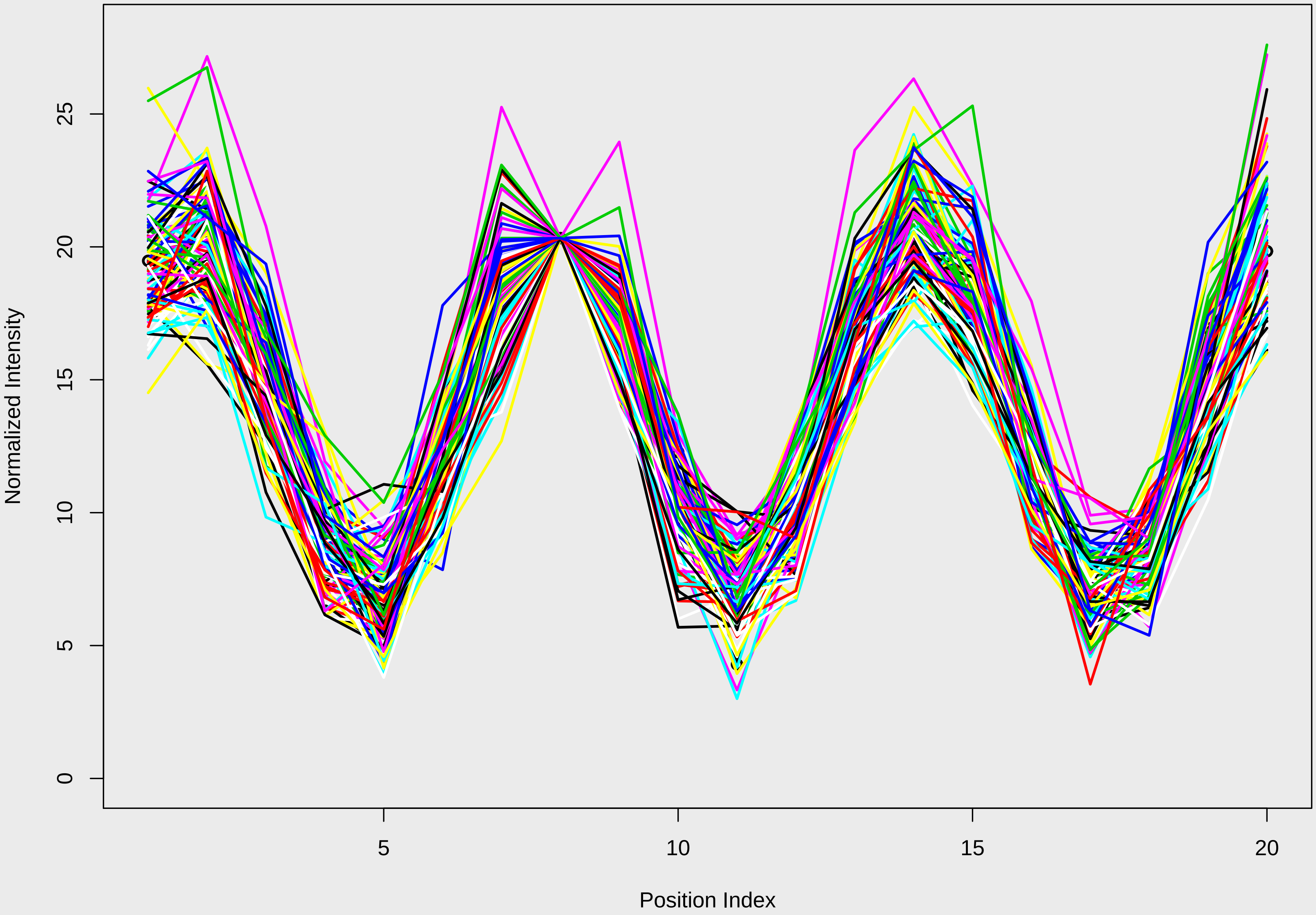
$$\begin{aligned} \boldsymbol{\epsilon}_i &\sim \text{N}_p(0, I_p) \;, \quad p = 20 \\ \alpha_i &\sim \text{log Normal}(0, 1) \end{aligned}$$

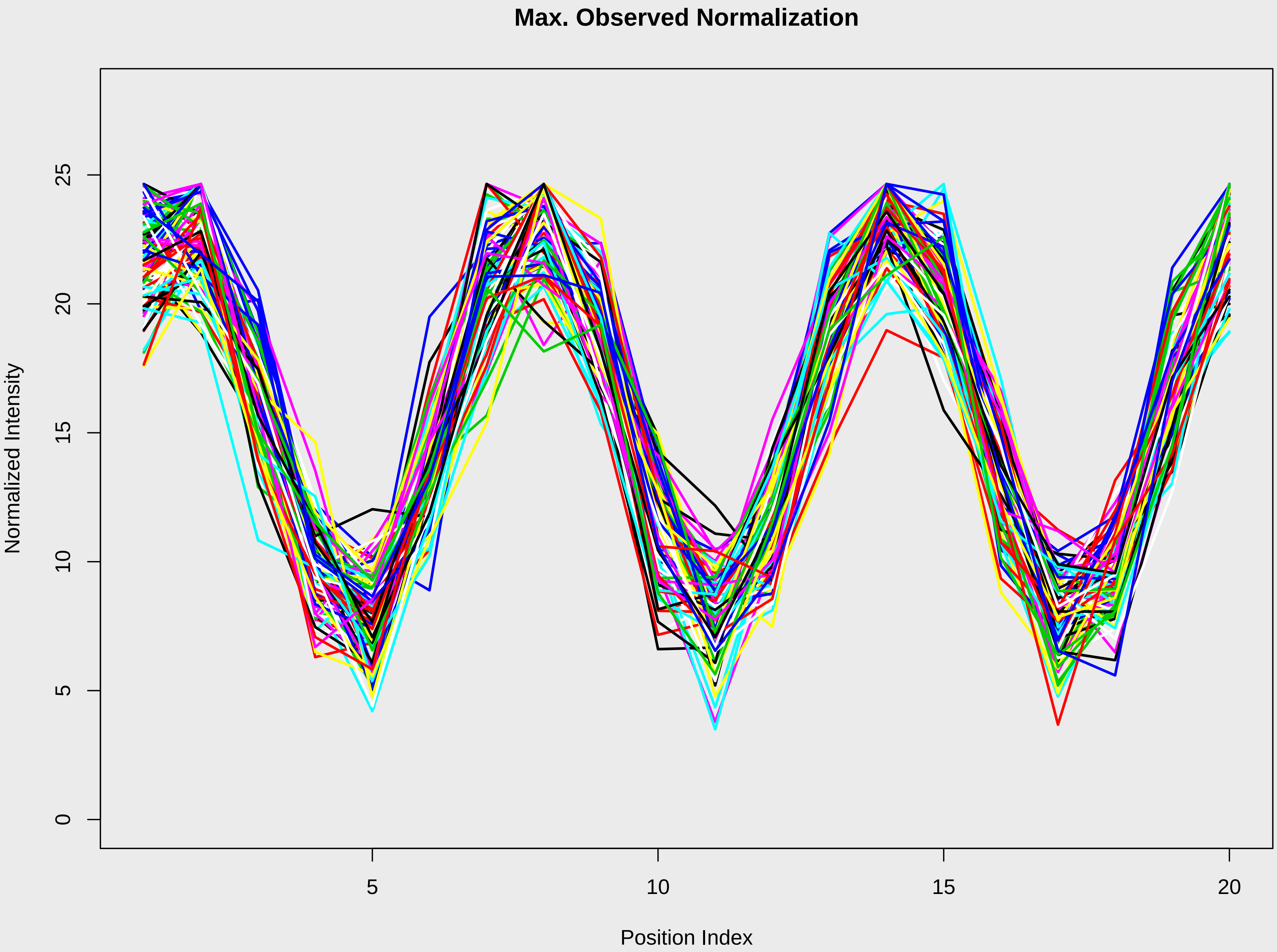
Evaluate 3 normalization methods

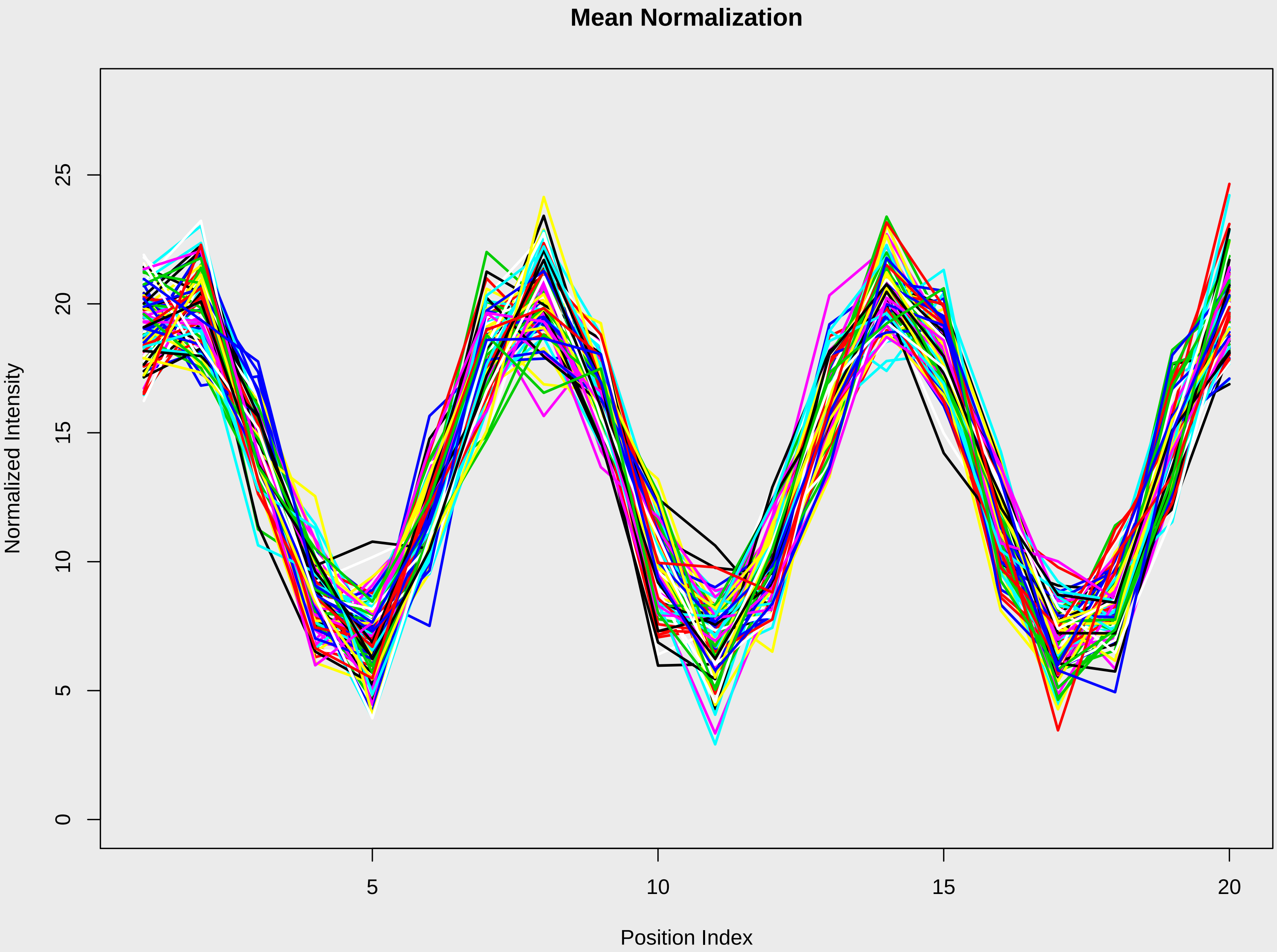
- normalization to a specified signal ($y_{i8} \approx \mu_8 = c$)
- maximum signal normalization $\max_t(y_{it}) = c$
- constant mean normalization $\mathbf{1}'\mathbf{y}_i = c$


Observed Frauda
Observed Intensity
0
20
40
60
80
5
10
15
20
Position Index

Component 8 Normalization




Max. Observed Normalization
Normalized Intensity
0
5
10
15
20
25
5
10
15
20
Position Index


Mean Normalization
Normalized Intensity
0
5
10
15
20
25
5
10
15
20
Position Index

Results of Frauda Normalization

Type of Transformation	Log-likelihood
Untransformed	-4016
Single Selected Peak ($t = 8$)	-2615
Observed Maximum	-2643
Mean Normalization	-2486

cDNA Microarray as an Example

- Many different methods of normalization used for cDNA microarray data.
  - Li and Wong (2001)
  - Schadt et. al. (2002)
  - Sidorov et. al. (2002)
  - Bolstad (2001)
- But, how does one choose between normalization methods? Evaluate different methods?

Currently, by examination of ratio versus intensity (RI) plots or other heuristic methods.

Model-based cDNA Microarray Data

Cui and colleagues (Statistical Applications in Genetics and Molecular Biology in 2003), describe the most ideal cDNA experiment, where Y_{ik} is the observed fuorescence intensity detected from both $i = r$ or g channels and $k = 1, \ldots, K$ spots.

That is,

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

Where the signal at “channel" i and gene k is comprised by the background signal, α_i, the concentration of the signal itensity, X_{ik}, and the slope of the linear relationship, β_i.

Model-based cDNA Microarray Data Simulation

But, for our simulation we assume that these values may have either multiplicative or additive errors and so the model is more realistically described by:

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

Where,

- $X_{ik} \sim$ lognormal (7, 1.1)
- Multiplicative errors:
  - $\eta_k \sim N$(0, σ_η^2)
  - $\zeta_{ik} \sim N$(0, $\sigma_{\zeta_i}^2$)
- Additive errors:
  - $\epsilon_k \sim N$(0, σ_ϵ^2)
  - $\delta_{ik} \sim N$(0,$\sigma_{\delta_i}^2$)

Data Simulation

For our examples, control of attributes and inducing of curvature into the RI plots is achieved by simulating data so that $\alpha_g \neq \alpha_r$, or $\beta_g \neq \beta_r$ where α_i controls background signal, and β_i are the channel slope values. We do not vary the error components to induce any distortion based

on multiplicative or additive error.

Data Transformations and Calculations of the Jacobians

For illustrative purposes we have used a selection of transformations described in the paper by Cui, et al.:

- Shift transformation (also called shift-log, discussed by Kerr, 2002)
- Loess transformation (Yang, 2000)
- Arsinh transformation (Huber, 2002)
- Linlog transformation (Cui, 2003)
- Linlog shift transformation (combination of the Shift and Linlog transformation)

In this talk, I will only show the form of two of these transformations, along with their Jacobians,for the sake of brevity. However, results will include all of these transformations.

Shift Transformation

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

The value of the constant, C, is calculated from minimizing the absolute deviations from each $\log_2$ ratio from the median ratio of the entire array.

The Jacobian for this transformation is:

$$\begin{aligned} \frac{\partial Z_{rk}}{\partial Y_{rk}} &= \frac{1}{(Y_{rk} - C)\log(2)} \\ \frac{\partial Z_{rk}}{\partial Y_{rk}} &= \frac{1}{(Y_{gk} + C)\log(2)} \end{aligned}$$

Loess Transformation

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

Where the constant, C_k, is a gene-specific constant obtained from the lowess fit.

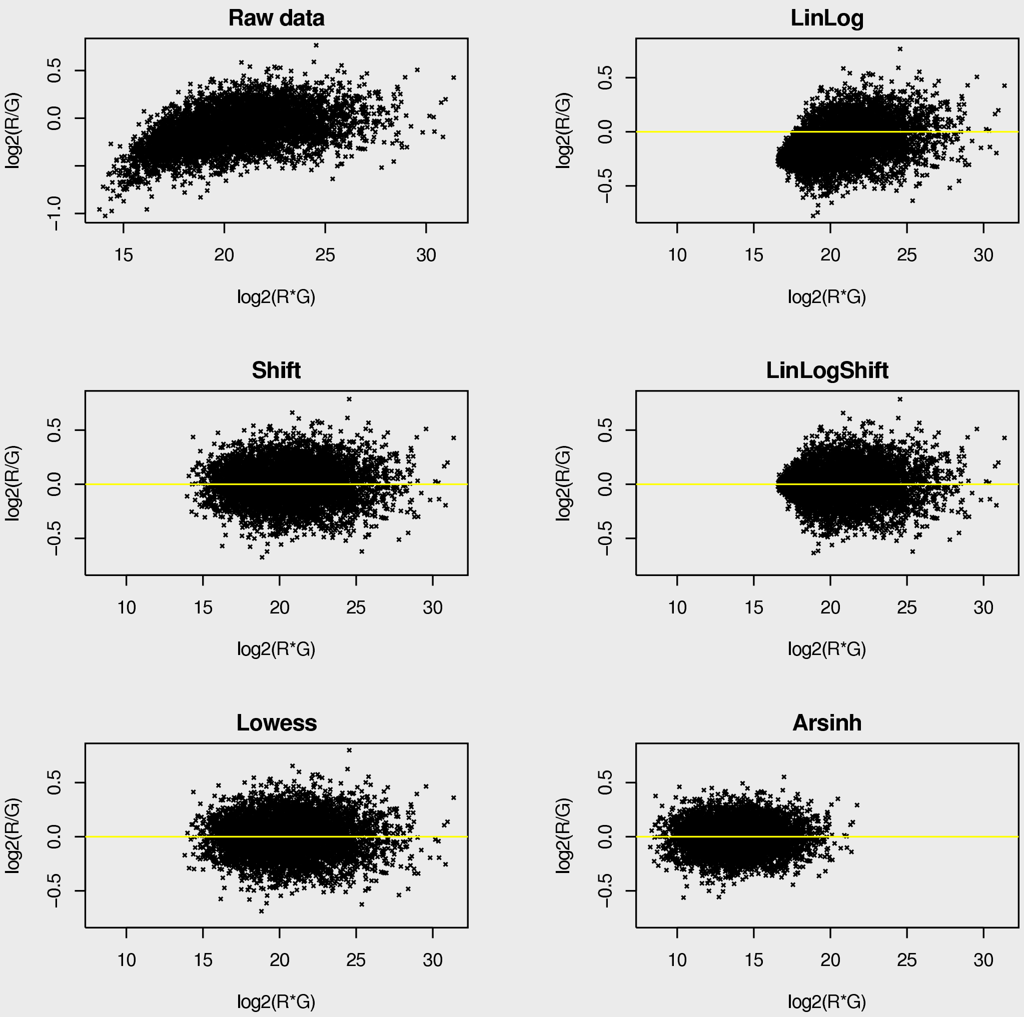
The Jacobian assumes C_k is constant and can be written as:

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

Background Differences

- α_r= 80
- α_g= 150
- β_r =β_g = 1

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


Raw data
log2(R/G)
0.5
0.0
−1.0
15
20
25
30
log2(R*G)
LinLog
log2(R/G)
0.5
0.0
−0.5
10
15
20
25
30
log2(R*G)
Shift
log2(R/G)
0.5
0.0
−0.5
10
15
20
25
30
log2(R*G)
LinLogShift
log2(R/G)
0.5
0.0
−0.5
10
15
20
25
30
log2(R*G)
Lowess
log2(R/G)
0.5
0.0
−0.5
10
15
20
25
30
log2(R*G)
Arsinh
log2(R/G)
0.5
0.0
−0.5
10
15
20
25
30
log2(R*G)

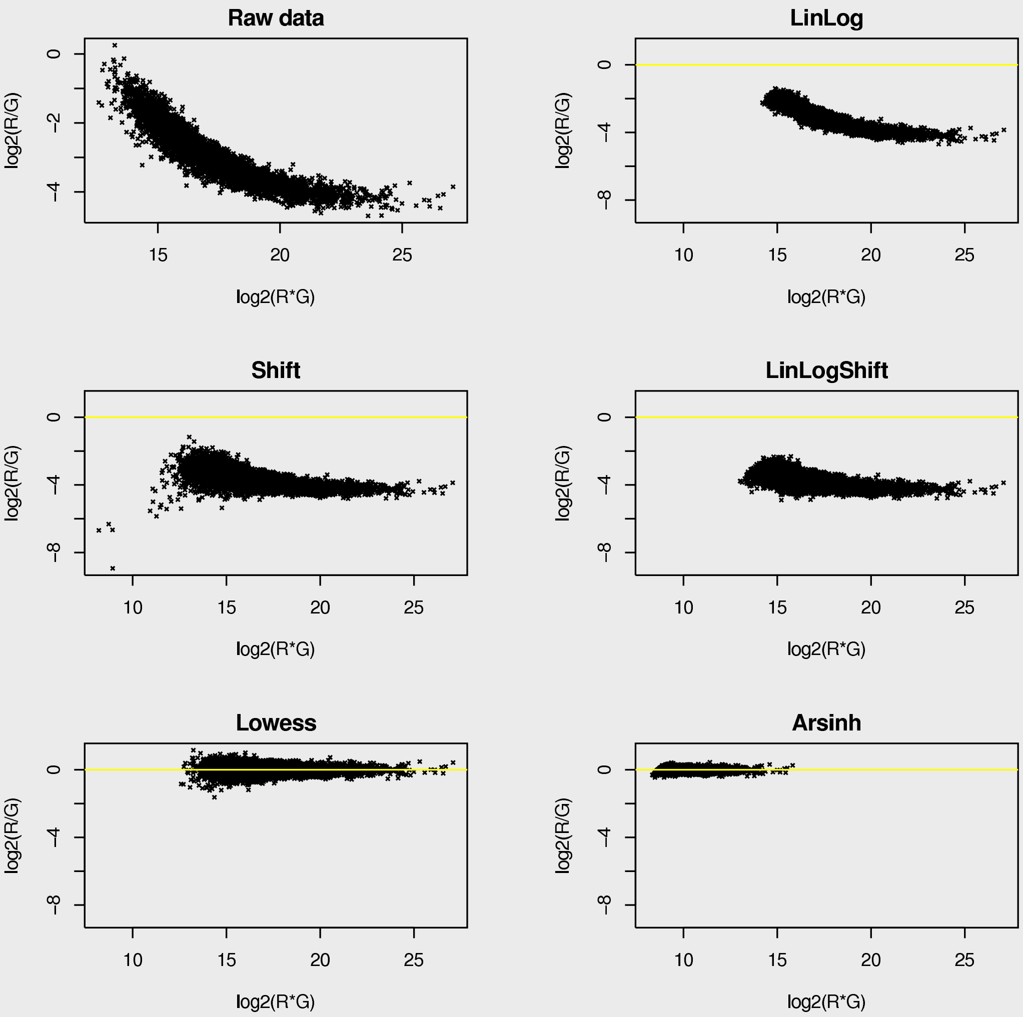
Likelihood Metrics for Background Differences

Transformation	Y Log-Likelihood	AIC	BIC
Shift	−66732	−66733	**−66736**
Lowess	**−66708**	**−66725**	−66786
LinLog	−69572	−69574	−69582
LinLogShift	−67552	−67555	−67566
Arsinh	−67976	−67980	−67994

Slope Differences

- β_r= 0.05
- β_g= 1
- α_r =α_g = 80

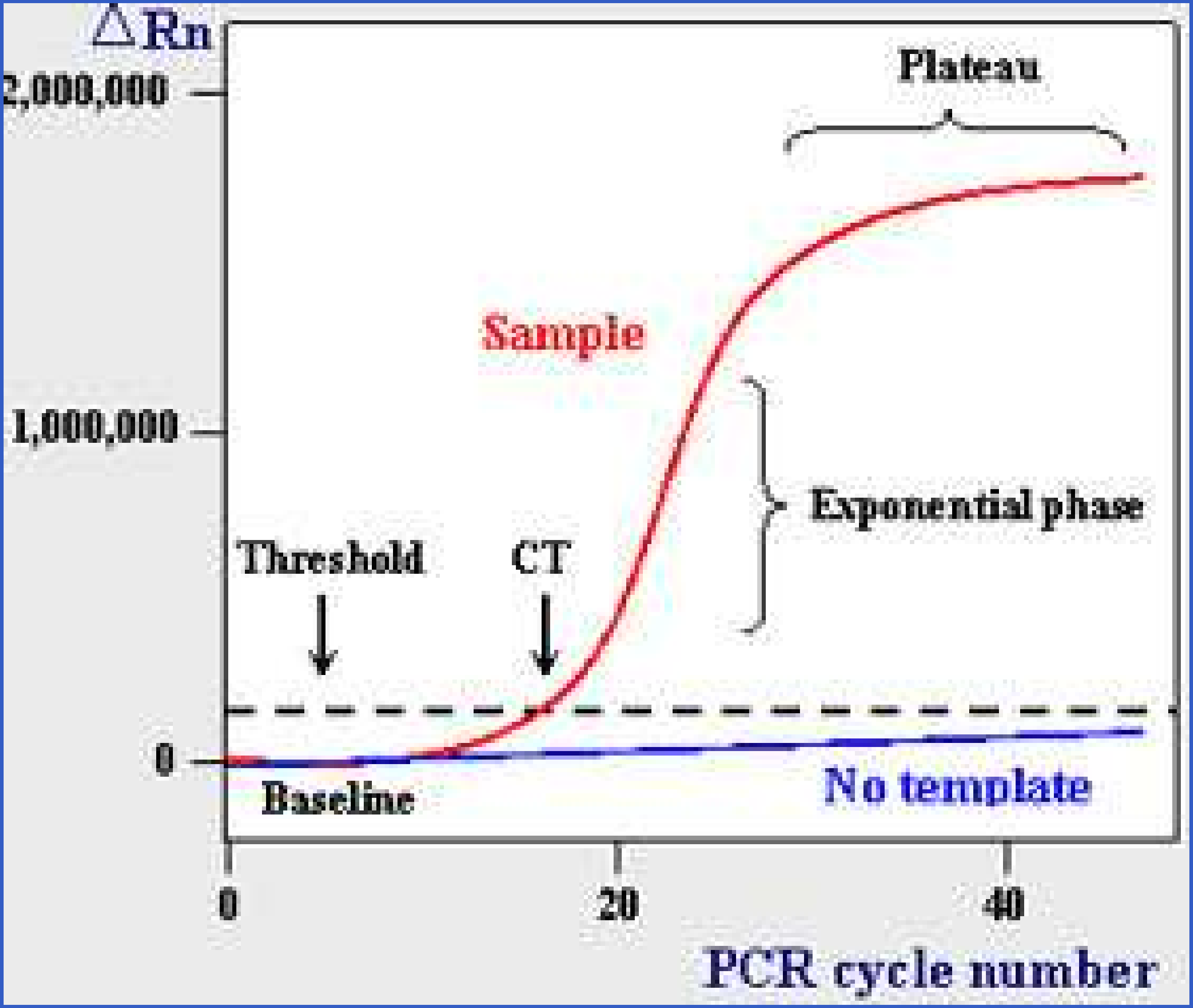
$$Y_{ik} = \alpha_i + \beta_i X_{ik} e^{\eta_k + \zeta_{ik}} + \epsilon_k + \delta_{ik}$$


Raw data
log2(R/G)
0
−2
−4
15
20
25
log2(R*G)
LinLog
log2(R/G)
0
−4
−8
10
15
20
25
log2(R*G)
Shift
log2(R/G)
0
−4
−8
10
15
20
25
log2(R*G)
LinLogShift
log2(R/G)
0
−4
−8
10
15
20
25
log2(R*G)
Lowess
log2(R/G)
0
−4
−8
10
15
20
25
log2(R*G)
Arsinh
log2(R/G)
0
−4
−8
10
15
20
25
log2(R*G)

Likelihood Metrics for Slope Differences

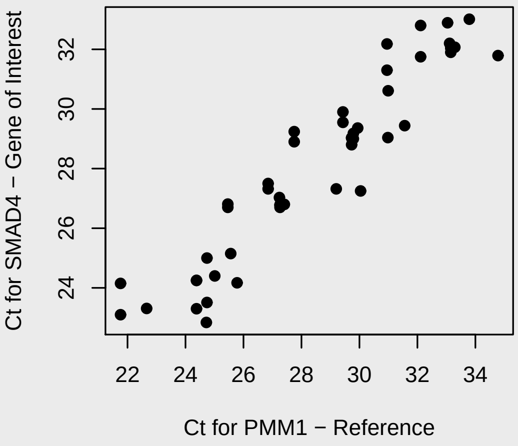
Transformation	Y Log-Likelihood	AIC	BIC
Shift	−72288	−72289	−72293
Lowess	**−57556**	**−57573**	**−57635**
LinLog	−74271	−74273	−74280
LinLogShift	−73631	−73634	−73645
Arsinh	−62539	−62543	−62558

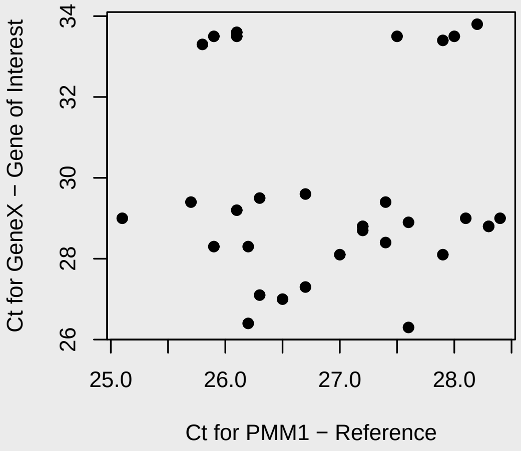
Real-Time RT-PCR

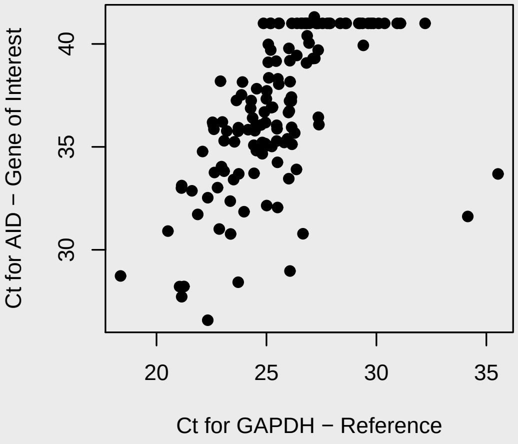


What Does "Normalization" Mean Here?

- Amount of starting product is unknown
- Relative quantitation


Ct for SMAD4 − Gene of Interest
24
26
28
30
32
22
24
26
28
30
32
34
Ct for PMM1 − Reference


Ct for GeneX − Gene of Interest
26
28
30
32
34
25.0
26.0
27.0
28.0
Ct for PMM1 − Reference


Ct for AID − Gene of Interest
30
35
40
20
25
30
35
Ct for GAPDH − Reference

Comparative Normalization Methods

- Difference in C_t values between gene of interest and reference, called the ΔC_t method

  (for many experiments this difference is then subtracted from another “plate” reference this is called the $\Delta\Delta C_t$ method)

- Model C_t for the gene of interest while “adjusting” for the C_t for the reference gene

In statistical terms, this would be comparing “models” with a parameter for the C_t value for the reference gene that is fixed at 1 or estimated.

Likelihood Metrics

	Log-Likelihood		AIC	
Transcript	Difference	Adjusted	Difference	Adjusted
SMAD4	305.2	286.9	309.2	292.9
AID	731.4	673.1	735.4	679.1
Gene X	149.6	145.4	153.6	151.4

Summary

- Statistical normalization is a data transformation to accommodate nuisance variation.
- Normalization constraints identify the statistical model, and induce singularities into the variance–covariance structure.
- Likelihood based metric gives a way to choose between competing methods.
- This statistical framework leads to principled development and evaluation of normalization methods.
- Good normalization is key to quantitative analysis.

Further Considerations

- Most problems require more complicated normalization schemes (multiple constraints, multiple parameters).
- When comparing transformations with differing numbers of constraints/parameters, a complexity penalty is required (BIC).
- There are links between normalization and errors–in–variables problems.
- Current stepwise “plug-in” estimation
    - each step depends on all preceding steps
    - any error is propagated forward
    - any uncertainty is ignored
    - instead use “simultaneous” modeling

Further Considerations – cont’d

- We are currently addressing specific issues of normalization in
  - Mass spec, other spectrometry methods (Dean)
  - Microarrays, immunohistochemistry, Tissue Microarray (Bonnie)
  - Manuscript focus on microarray methods – resubmitted

`bonnie.lafleur@hsc.utah.edu`