Program #: 219
Board #: B82

Sometimes Simpler Is Better: VLog, An Easy-to-Implement Log-Like Transform for Cytometry

Bagwell CB[1], Hill BL[1], Hunsberger BC[1], Herbert DJ[1], Bray CM[1]; [1] Verity Software House, 45A Augusta Rd, Topsham, ME 04086 USA.

Abstract

Typically, the magnitude of a measurement's uncertainty is proportional to the magnitude of the signal. The proportionality constant that relates uncertainties to signal magnitudes is coefficient of variation or cv and is given by the simple equation, sd = cv*x where x is the magnitude of the measurement and sd is its uncertainty or standard deviation. Because biologic signals have ranges that span four to five decades, the uncertainties of population measurements also span many decades. If you wanted to design a seemingly perfect transformation for cytometry where the linearly increasing sd's would be converted to uniform sd's, you would find a function whose slope at each x value would be 1/x. This function is the well-known log function, z=log(x), and has been routinely employed for many years to enable the visualization of cytometrically-derived cellular populations.

The primary problem with log transforms is that they are not defined for signals that are less than or equal to zero due to the 1/x slope not being defined at x=0. Unfortunately, most signal un-mixing and base-line restore algorithms can create zero or less than zero numbers. Truncating these negative signals to the first channel of display systems has created numerous issues for cytometry - the worst being a strong tendency to over compensate data. In 2002 Parks and Moore described a generalized hyperbolic sine function that was log-like at higher signal intensities and linear-like through the origin. Since then, there have been a number of other log-like transforms that have been published or described in earlier literature. Many of these transforms are quite complex and present challenges to software implementers.

This study examines a simple log-like transform that is very easy to implement and that has the desirable attributes of being log-like at higher intensities and linear-like through the origin. The arguments to the transform are 1) number of desired decades, 2) a scale-independent coefficient that determines the slope through zero, 3) desired maximum transform range, and 4) the maximum linear range. Implementers are free to use and modify this transform if they don't have access to other more complex transformation systems.

Introduction

The general topic of creating log-like transforms that admit zero or negative numbers has been well explored in other scientific disciplines since 1949 (1-11). In 2002 Parks et al. were the first cytometrists to investigate the general form of the hyperbolic sine function as a potential solution to the problem (12) and then later published their "Logicle" transform in 2006 (13) and revised it slightly in 2012 (14). HyperLog is also a log-like transform that accepts zero or negative valued numbers and was published in 2005 (15). Both transform implementations are functions that tend to be linear through the origin and logarithmic away from the origin. Although there has been general acceptance of log-like transforms and their application to cytometry data, the detailed implementations are often not trivial, many times involving numerical root finding routines.

A detailed analysis of cytometric measurement sources of variance is also well-described in the literature (16-21). There are three basic components of measurement variance: 1) gain-dependent variability, 2) photo-electron counting error, and 3) signal-independent sources of error. Gain-dependent variability has the general characteristic that measurement uncertainty is proportional to the gain applied to a specific signal. The gain can be either biologic or electronic in nature. The proportionality constant, coefficient of variation (cv), typically characterizes this type of variability.

For fluorescence-based cytometers, while a cell is bathed in laser light, fluorochromes, frequently attached to antibodies, will emit and re-emit fluorescence photons in all directions numerous times. A relatively small fraction of these photons are detected and converted to photo-electrons by detectors such as photo-multiplier tubes (pmts). The counting error associated with these captured photo-electrons is typically assumed to be Poisson-distributed.

The third source of measurement variability is from a number of sources. Some of this variability is due to the counting error associated with non-signal photo-electrons, which includes sources such as ambient light, light-scatter, and Raman scattering. However, the major sources of signal-independent variability are due to cellular autofluorescence, compensated secondary detector counting error, and the implicit wobble associated with base-line restore algorithms.

The purpose of this study is to present and examine a data transform that was designed to stabilize the variability from all three sources. This transform, VLog, not only has efficient closed form solutions, but also can be enhanced such that its parameters rarely have to be recomputed with different data sets.

Materials and Methods

Mathematical Data Simulations: Mathematical analysis and presentation graphics were done using Mathcad 15.0, Parametric Technology Corporation (PTC), Needham, MA, USA.

Normally distributed random numbers were generated with the Box-Muller equation,

$$Norm(\mu,\delta) = \mu + \delta\sqrt{-2\ln(rnd(1))}\cos(2\pi\cdot rnd(1)).$$

Gain-dependent Variances for Synthesized Populations:

The first four gain-dependent populations were synthesized as shown below,

$$X_{i,g} = Norm(\mu,\delta)\cdot G_g,$$
$$where,\ \mu = 50, \delta = 5,\ \ n = 1000,$$
$$i = 1,2,...,n;\ g = 1,2,...,4,$$
$$G_g = g.$$

The second four gain-dependent populations were synthesized as,

$$X_{i,4+g} = Norm(\mu,\delta)\cdot G_g,$$
$$where,\ \ g = 1,2,...,4,$$
$$G_g = 10^g.$$

These data were used in **Figure 1** to demonstrate the general utility of the log transform.

General Variances for Synthesized Populations:

The eight populations shown in **Figure 4** were synthesized as shown below,

$$X2_{i,g} = Norm\left(M_g, \delta\left(M_g, cv, q_c, b\right)\right),$$
$$where,\ \delta(x; cv, q_c, b) = \sqrt{cv^2x^2 + \frac{x}{q_c} + b^2},$$
$$g = 1,2,...,8,\ M^T = [10\ \ 50\ \ 150\ \ 400\ \ 1000\ \ 2800\ \ 8000\ \ 17000],$$
$$cv = 0.03, q_c = 0.6, n = 1000, b = 6.0, i = 1,2,...,n.$$

These data were used in **Figure 4** to demonstrate the incompleteness of the Log Transform and the efficacy of the VLog Transform.

Data Sets: the files represented in **Figure 6** mainly come from a repository of files from a published study (22) and from Helios data described in another presentation (23).

1

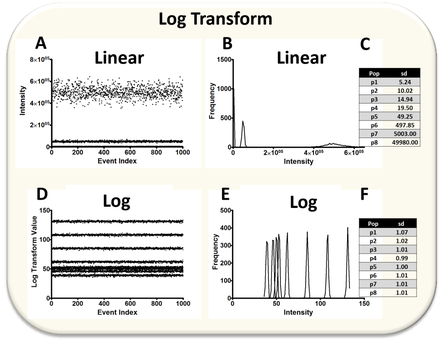

Figure 1. Log Transform: The longstanding utility of the log transform is primarily due to its ability to stabilize a set of population gain-dependent variabilities. The increased dynamic range associated with the transform is really a useful side-effect of this stabilizing capability. If the variability of measurements is only determined by the equation,

$$\delta(x; cv) = cv\cdot x. \qquad Eq.1$$
$$where,\ \delta = \text{standard deviation},$$
$$cv = \text{coefficient of variation}.$$

then the log function would be the transform of choice since,

$$z(x; cv) = \int_1^x \frac{1}{cv\cdot t}\,dt = \frac{1}{cv}\cdot\ln(x). \qquad Eq.2$$

Panels A, B, and C show the linear data from populations with standard deviations (sd's) determined solely by **Eq.1** (see **Gain-dependent populations in M&M**). **Panel A** shows each event's intensity value plotted against event index and **Panel B** shows the associated frequency histograms. Because of the linear increasing sd's over a large dynamic range (see **Panel C)** only two populations are adequately represented in the linear domain.

Panels D, E, and F show the log transformed data using **Eq. 2**. Each transformed sd is relatively uniform with values very close to unity (see **Panel F**).

2

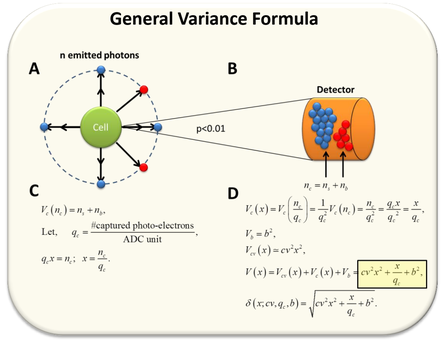

Figure 2. General Variance Formula: For fluorescence-based cytometers, while a cell is bathed in laser light, fluorochromes emit and re-emit fluorescence photons in all directions numerous times (10^3-10^9, see **A**). A relatively small fraction of these photons (e.g. 0.01) ultimately are captured and typically produce a specific number of photo-electrons at the primary detector (see **B**). Some of these captured photo-electrons are from the fluorochrome of interest (n_s, see blue spheres) and some are not (n_b, see red spheres). The non-signal photo-electrons are generated from a wide variety of sources, where the major source is usually the particle's background fluorescence. Generally, these non-signal photons are grouped together and symbolized as n_b (see **B** and **C**).

The variance of the number of captured photon-electrons, $V_c(n_c)$ is predicted by the Poisson Distribution as equal to $n_s + n_b$ (see **C**). These captured photo-electrons will eventually be amplified and digitized to form x ADC units. If this process is linear, then there will be a proportionality constant, q_c, that can convert number of captured photo-electrons to x ADC units (see **C**). By substitution, the counting variance of x is therefore given as $V_c(n_c/q_c)$ which can be simplified to x/q_c (see **D**). There are also a number of x-independent sources of variability that can be grouped together as $V_b = b^2$.

The major source of variability in measurement systems like cytometry are gain-dependent (biological and electronic) and are generally characterized by the coefficient of variation or cv. The variance due to cv, V_{cv}, is approximately equal to cv^2*x^2 (see **D**). The total variance associated with x is the summation of these three variances and is given by formula shown in **panel D**'s black box with yellow highlight. Note that the cv variance increases with the square of x and therefore is the dominant variance, the count-dependent variance increases linearly with x, and the background variance is independent of x. The standard deviation or sigma function is the square-root of the variance (see bottom of **D**).

3

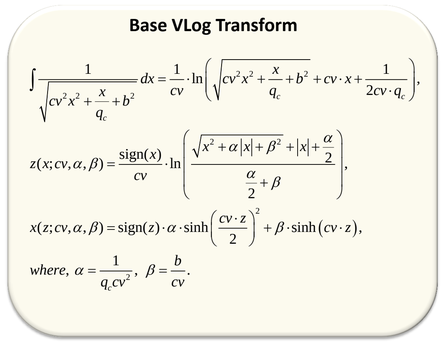

Figure3. Base Vlog Transform: Measurement uncertainties that follow the general variance formula (see **Figure 2**) can be stabilized by finding the integral of,

$$\frac{1}{\delta(x; cv, q_c, b)} = \frac{1}{\sqrt{cv^2x^2 + \frac{x}{q_c} + b^2}}.$$

As shown in **Figure 3**, the solution is in closed form. The integral formula is then re-parameterized to eliminate cv from the ln function argument, translocated to z=0 at x=0, and then made symmetric about the z=0 axis. The shape-dependent parameters are α and β. Interestingly, when β>0 and α=0, the transform behaves as the hyperbolic sine.

4

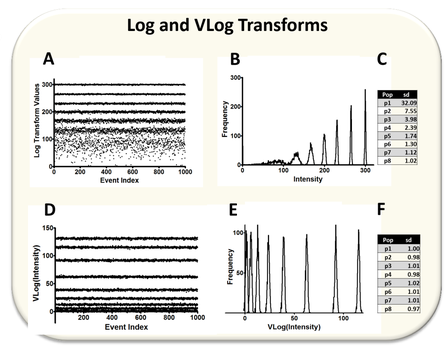

Figure 4. Log and VLog Transforms: **Panels A, B, and C** show the Log Transform of the linear data from populations with standard deviations (sd's) determined by the general variance formula (see **General Variances for Synthesized Populations in M&M**). **Panel A** shows each event's Log transformed intensity value plotted against event index and **Panel B** shows the associated frequency histograms. Because the Log Transform only stabilizes the gain-dependent portion of the total variance, the sd's increase with lower intensity values (see **Panel C** for enumerated sd's).

Panels D, E, and F show the transformed data using the VLog . Each transformed sd is relatively uniform with values very close to unity (see **Panel F**).

5

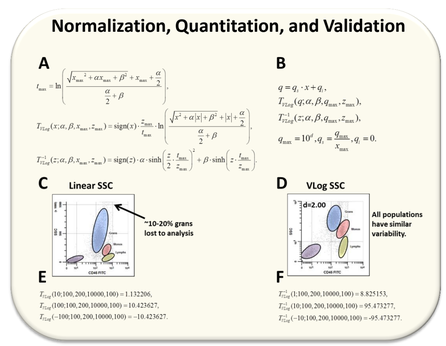

Figure 5. Normalization, Quantitation, and Validation: **Panel A** shows the formulation of VLog that is normalized such that at x=x_{max}, the transform is at z=z_{max} and at x=$-x_{max}$, z=$-z_{max}$. The variable t_{max} only needs to be evaluated once for a specific set of parameters. **Panel B** shows the additional changes necessary to support quantitative axes. The bottom of **Panel B** demonstrates how to standardize the transform to a specific number of decades, d, that better reflects the true dynamic range in the data. **Panels C and D** demonstrate the important application of defining a particular number of decades to light-scatter measurements that typically contain a low number of decades of information. **Panels E and F** show some validation equations for those interested in implementing the VLog transforms.

6

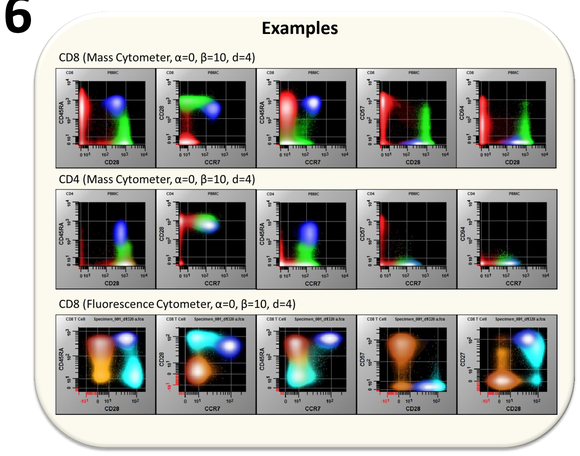

Figure 6. Examples: The example surface plots shown in this figure come from a repository of files from a published study (22) and from Helios data described in another presentation (23). Colors are blended from probabilistically defined stages during GemStone modeling . All the transforms used had the same parameters: α=0, β=10, d=4. In some cases only a portion of the scale is showing due to automatic zooming logic.

Discussion

Log and Log-Like transforms for cytometry are variable-sloped functions that tend to stabilize variances to better enable population visualization and analysis. Most all these variances are described by a general variance formula (see **Figure 2**) that blends gain-dependent, counting-error, and signal-independent variabilities. The log transform only partially stabilizes the gain-dependent variances, which is why cytometrically-derived population variances tend to increase with lower intensities. Also, because the log transform is not defined for zero and less than zero values, serious distortions can occur near the origin that can lead to inappropriate conclusions about the true nature of the data.

By numerically integrating the general variance formula's reciprocal, a transform called VLog can be derived and mathematically represented in efficient closed-form equations (see **Figures 3 and 5**). The relative simplicity of these equations along with their generality potentially make VLog a useful tool for cytometry and possibly other technologies. Software engineers interested in exploring VLog's capabilities are free to do so. Validation equations have been added in **Figure 6** to help with these implementations.

Enhancing the transform to include the capability of supporting quantitative axes also solves an increasingly problematic issue with cytometry displays. As the ADC maximum ranges have increased over time, many cytometer display systems have also increased the number of decades displayed on their axes. Rather than considering the ADC max range as the source for number of decades, the quantitative system enables the user to approximate the number of measureable decades of information encoded in the data. For immunofluorescence measurements, typically four to five decades are all that are needed; however, for light-scatter measurements only 1.5 to 2 decades are normally required. By matching the number of decades to the biology rather than the electronics, different cytometers with different ADC max ranges can be made to produce similar distributions. **Figure 6** amplifies this point by demonstrating that different markers and different types of cytometers can produce very similar data patterns with exactly the same transform setup parameters.

References

1. Johnson NL. Systems of frequency curves generated by methods of translation. Biometrika 1949;36:149-176.
2. Bickle PJ, Doksum KA. An analysis of transformations revisited. J Amer Statist Assoc 1981;76:296-311.
3. Box GEP, Cox DR. An analysis of transformations. J Royal Statist Soc. 1964;26:211-243.
4. Burbidge JB, Magree L, Robb AL. Alternative transformations to handle extreme values of the dependent variable. J Amer Statist Assoc 1988;83:123-127.
5. Layton DF. Alternative approaches for modeling concave willingness to pay functions in conjoint valuation. Amer. J. Agr. Econ 2001;83:1314-1320.
6. Guarnieri MD, Ortolani S, Montegriffo P, Renzini A, Barbuy B, Bica E, Moneti A. Infrared array photometry of bulge globular clusters. Astron Astrophys 1998;331:70-80.
7. Munson PA. Consistency' test for determining the significance of gene expression changes on replicate samples and two convenient variance-stabilizing transformations. GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data 2001.
8. Tukey JW. On the comparative anatomy of transformations. Ann of Math Stat 1964;28:602-632.
9. Tukey JW. Exploratory data analysis: Addison-Wesley, Reading, MA; 1977.
10. Holder D, Raubertas RF, Pikounis VB, Svetnik V, Soper K. Statistical analysis of high density oligonucleotide arrays: a safer approach. 2001.
11. Rocke DM, Durbin B. Approximate variance-stabilizing transformations for gene-expression microarray data. Bioinformatics 2003;19:966-972.
12. Parks DR, Moore W. 2002 Oct 18-21; Asilomar Conference Center in Pacific Grove, California.
13. Parks DR, Roederer M, Moore WA. A new "Logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. Cytometry A 2006;69:541-51.
14. Moore WA, Parks DR. Update for the logicle data scale including operational code implementations. Cytometry A 2012;81:273-7.
15. Bagwell C. Hyperlog-a flexible log-like transform for negative, zero, and positive valued data. Cytometry A 2005;64:34-42.
16. Steen HB. Noise, Sensitivity, and Resolution of Flow Cytometers. Cytometry 1992;13:822-830.
17. Wood JC, Hoffman RA. Evaluating fluorescence sensitivity on flow cytometers: an overview. Cytometry 1998;33:256-9.
18. Hoffman R, Wood C. Characterization of Flow Cytometer Instrument Sensitivity. Current Protocols in Cytometry. Volume 1.20: John Wiley & Sons, Inc.; 2007. p 1-18.
19. Wood JC. Flow cytometer performance: fluorochrome dependent sensitivity and instrument configuration. Cytometry 1995;22:331-2.
20. Wood JC. Fundamental flow cytometer properties governing sensitivity and resolution. Cytometry 1998;33:260-6.
21. Perfetto SP, Chattopadhyay PK, Wood J, Nguyen R, Ambrozak D, Hill JP, Roederer M. Q and B values are critical measurements required for inter-instrument standardization and development of multicolor flow cytometry staining panels. Cytometry A 2014;85:1037-48.
22. Inokuma MS, Maino VC, Bagwell CB. Probability state modeling of memory CD8(+) T-cell differentiation. J Immunol Methods 2013;397:8-17.
23. Bagwell C, Hill B, Wood B, Wallace P, Alrazzak M, Kelliher A, Preffer F. Human B-Cell and Progenitor Stages As Determined by Probability State Modeling of Multidimensional Cytometry Data. Cytometry B Clin Cytom 2015.
24. Bagwell CB, Leipold M, Maecker H, Stelzer G. High-dimensional modeling of peripheral blood mononuclear cells from a Helios Instrument. 2016 June 11-15, 2016; Washington State Convention Center, Seattle, Washington, USA.