Variance Stabilizing Transformations for image-based compound profiling features

LEonard Wafula

August 26, 2017

Introduction

Image-based multi-parameteric compound profiling features

□ Biological method used as a proxy for distinguishing compounds in the drug discovery chain using a range of features extracted from image-based assays by applying High Throughput Microscopy (HTM)
□ The features provide information on

i. Intracellular biomarkers: texture, intensity, spatial distribution etc
ii. Cells: shape, geometry, quantity .. .

□ Why

i. understand how compounds induce their desired properties and describe their mechanisms of action
ii. preferentially identify highly specific compounds having a desired effect on a given biological target
iii. early detection of undesired compound effects on cells + cellular activity: toxicity

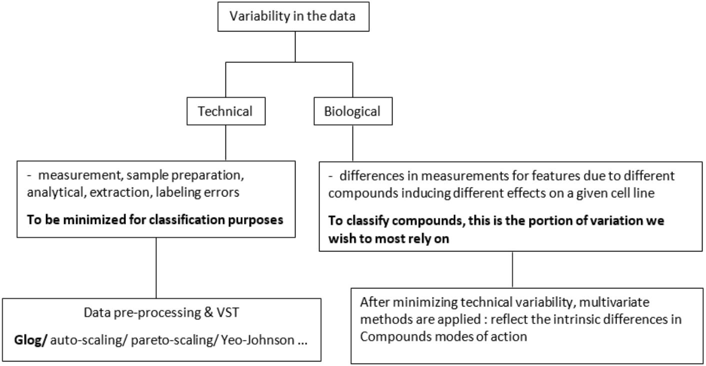
.. introduction

It's all good, but .... Most of these features often

□ are highly correlated: need to limit features used for analysis

□ have non-normal distributions: mean-variance relationship present

- multivariate classification methods hugely depend on variance



Variability in the data

Technical

Biological

- measurement, sample preparation, analytical, extraction, labeling errors

To be minimized for classification purposes

- differences in measurements for features due to different compounds inducing different effects on a given cell line

To classify compounds, this is the portion of variation we wish to most rely on

Data pre-processing & VST

Glog/ auto-scaling/ pareto-scaling/ Yeo-Johnson ...

After minimizing technical variability, multivariate methods are applied : reflect the intrinsic differences in Compounds modes of action

Aim of the analysis

To assess the:

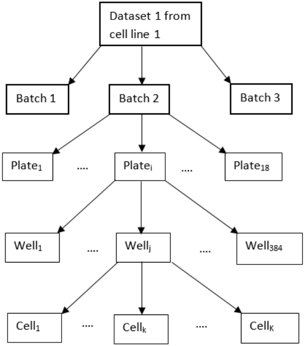
I. effect of glog transformation on separation of treatment replicates from non-replicates

II. effect of glog transformation on proportion of actively-called treatments

III. performance of a glog transformation on treatments separation when applied at cell- or well-level

♣ Treatment: a compound at a given concentration (a compound can have 4 or 5 concentration levels - 1 µM (microMolar), 3 µM, 3.34 µM, 9 µM and 11.1 µM)

Data, the glog transformation and data pre-processing



Data

i. 2 cancer cell lines: Liver & Colon

ii. a plate had btwn 134909 & 281177 (152679 & 330117) in $1^{st}(2^{nd})$ data

iii. a well had btwn 73 & 1812 (95 & 2120) cells in the $1^{st}(2^{nd})$ data

iv. 311 compounds including DMSO control

v. we tested a total of 1253 treatments

vi. 462 features extracted from each cell

..Data, the glog transformation and data pre-processing

Glog transformation

- formula

$$z = \text{Log}(y - \alpha + \sqrt{(y - \alpha)^2 + \lambda})$$

- where
  - z: glog-transformed data
  - y: untransformed data
  - α: feature mean across DMSO controls
  - λ: transformation parameter

Data pre-processing

- Aggregation - calculating mean for each feature per well
- Normalization -

$$\frac{\text{feature}_{value} - \text{mean.feature}_{DMSO}}{\text{pooled.SD.feature}_{across.plates}}$$

- Feature selection
  - MRMR: identify set of features with low pairwise correlation & high reproducibility among replicates.
  - AUC value for btwn 2-75 features
  - optimal feature: maximizes separation of treatment replicates within 1 Std error of AUC
- Active calling: treatments with $\geq 50\%$ active replicates

Methodology

⋆ Hotelling's T^2 method

- measures difference in 2 multivariate means
- formula

$$T^2 = \frac{(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)\prime(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)}{\boldsymbol{S}_p(\frac{1}{n_1} + \frac{1}{n_2})}$$

- normality assumptions for optimal results
- Only actively-called treatments in pre- & post-transformation used
- $+$ shift in T^2 distribution indicate improved treatment separation

⋆ AUC method

2-steps involved in AUC-calculation

- Pearson correlation btwn pairs of replicates (& non-replicates) were calculated & distributions plotted
- separation btwn the 2 distribution quantified by constructing an ROC curve using a series of correlation thresholds & calculating an AUC value
- transformations leading to higher AUC values ⇢ improved treatments separation compared to corresponding untransformed data

Results: EDA

No. of replicates(% of sample treatments)		
9 (%)	36 (%)	45 (%)
1192(95.208)	28(2.236)	32(2.556)

- DMSO control replicated across 1512 wells
- For both data sets
- Implications
  - For calculation of Hotelling's T^2, a limited number of selected features was used to maximize its power
  - 10 highest ranked features from MRMR used to calculate T^2

Transformations effects on treatments separation

Prologue

◇ Only glog transformations of λ equal to 0.1 and, 0.5 to 25 at 0.5 interval investigated for both T^2 and AUC methods

◇ Each transformed data compared to its corresponding untransformed data defined by actively-called treatments present both pre- and post-transformation

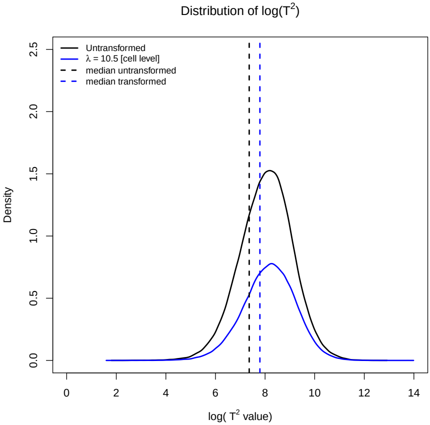
◇ Improved treatments separation shown by $+$ve shifts in distribution (and/or associated statistics) of T^2 for transformed compared to untransformed, and/or higher AUC values

◇ Results presented for the first cell line only since results largely led to similar conclusions for both cell lines

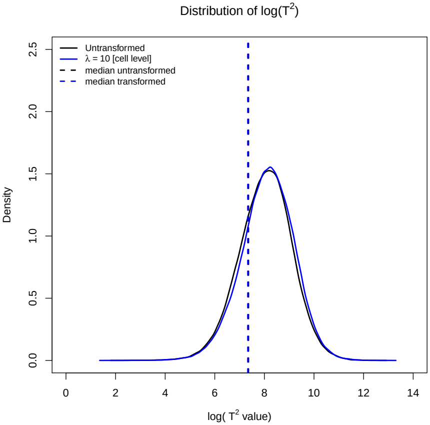
Transformations effects on treatments separation - T^2

※ Presence of very high [very different] and very low [highly similar] values

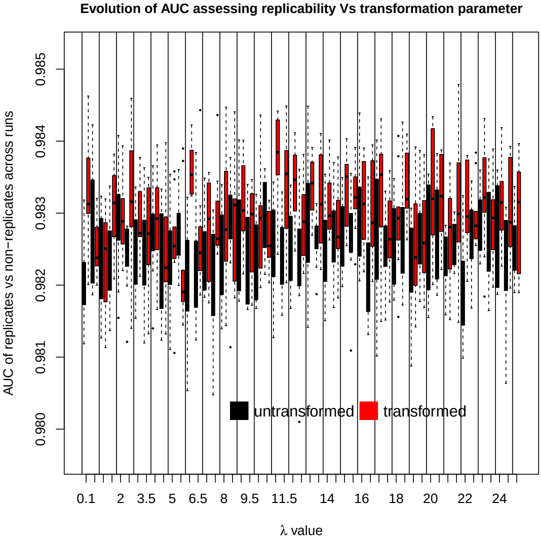
※ Some led to slight but negligible improvements (e.g $\lambda = 10.5$)



※ Others led to no improvements (e.g $\lambda = 10$)



Transformations effects on treatments separation - AUC



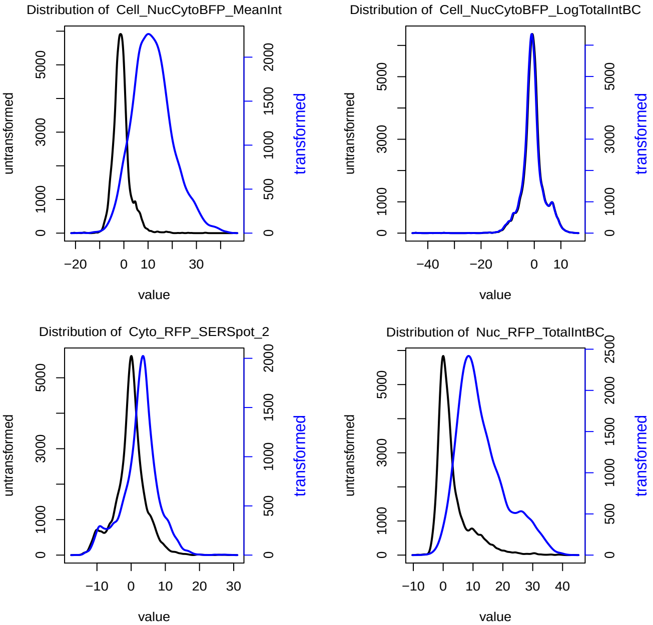
- high AUC b4-transformation
- some (e.g $\lambda = 0.1$) led to marginal increases
- others (e.g $\lambda = 5.5$) separated slightly poorer
- the differences were however very minimal & non-significant

Transformations effects on treatments separation- Epilogue

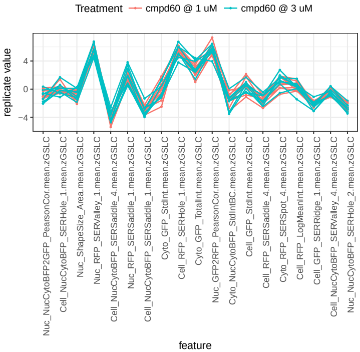
⋄ In both methods, minimal & insignificant differences were observed: Transformations failed to improve treatments separation

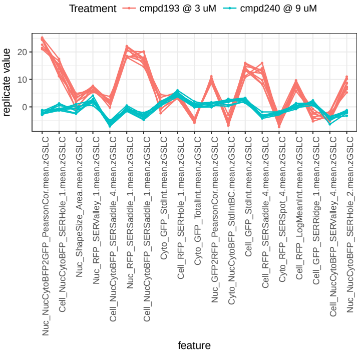
⋄ Why?

1. Transformation effect on features distributions

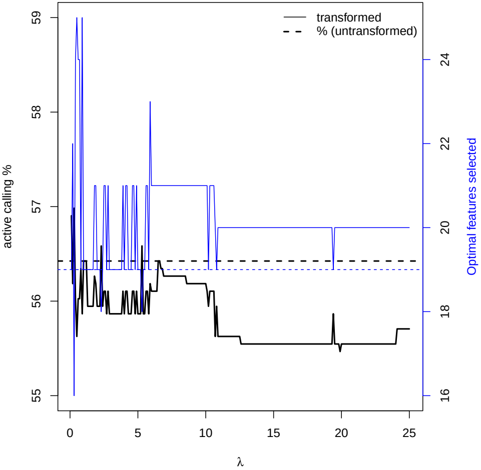


2. Differentiating ability of features selected (before transformation)



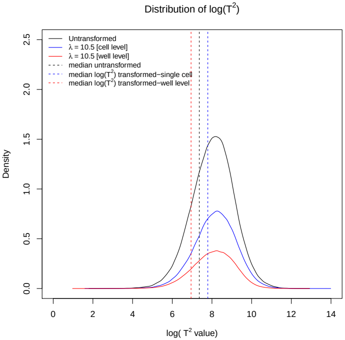


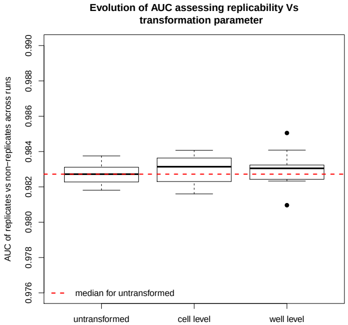
Effect of transformation on treatments active-calling



- lower % of active-calling
- intriguing relationship btwn number of features selected & prop. of active-calling
- similar relationship in 2^{nd} cell line

Transform at cell- or well-level? [T^2 & AUC approaches]





$\sim$ For $\lambda = 10.5$

$\sim$ Minimal non-significant improvement in treatment separation when transformed at cell-level $>$ well level

$\sim$ High AUC values pre-transformation

$\sim$ No clear preference for cell- or well-level transformation

Discussion

From our study, we observed that:

$\sim$ Transformations did not improve treatments separation beyond what was seen pre-transformation

$\sim$ Transformations led to lower(higher) proportion of active-calling in 1^{st}(2^{nd}) data

$\sim$ Inverse relationship between proportion of active-calling and number of features selected was evident

$\sim$ There was no preference in transforming data at cell- or well-level