Published in final edited form as:

J Comput Graph Stat. 2020 ; 29(4): 929–941. doi:10.1080/10618600.2020.1741379.

Automatic Transformation and Integration to Improve Visualization and Discovery of Latent Effects in Imaging Data

Gregory J. Hunt,

Department of Mathematics, William & Mary

Mark A. Dane,

Department of Biomedical Engineering, Knight Cancer Institute, OHSU Center for Spatial Systems Biomedicine, Oregon Health and Science University

James E. Korkola,

Department of Biomedical Engineering, Knight Cancer Institute, OHSU Center for Spatial Systems Biomedicine, Oregon Health and Science University

Laura M. Heiser,

Department of Biomedical Engineering, Knight Cancer Institute, OHSU Center for Spatial Systems Biomedicine, Oregon Health and Science University

Johann A. Gagnon-Bartsch

Department of Statistics, University of Michigan

Abstract

Proper data transformation is an essential part of analysis. Choosing appropriate transformations for variables can enhance visualization, improve efficacy of analytical methods, and increase data interpretability. However determining appropriate transformations of variables from high-content imaging data poses new challenges. Imaging data produces hundreds of covariates from each of thousands of images in a corpus. Each of these covariates will have a different distribution and need a potentially different transformation. As such imaging data produces hundreds of covariates, determining an appropriate transformation for each of them is infeasible by hand. In this paper we explore simple, robust, and automatic transformations of high-content image data. A central application of our work is to microenvironment microarray bio-imaging data from the NIH LINCS program. We show that our robust transformations enhance visualization and improve the discovery of substantively relevant latent effects. These transformations enhance analysis of image features individually and also improve data integration approaches when combining together multiple features. We anticipate that the advantages of this work will likely also be realized in the analysis of data from other high-content and highly-multiplexed technologies like Cell Painting or Cyclic Immunofluorescence. Software and further analysis can be found at gjhunt.github.io/r/

Keywords

imaging; latent variables; visualization; automatic transformation; data integration; PCA

1 Introduction

Transformation of data is an essential component in many areas of analysis. Consider principal components analysis (PCA), one of the primary statistical techniques for visualization and recovery of latent variables. PCA is well-known to be sensitive to skewed distributions and outliers (Hubert et al., 2009; Maadooliat et al., 2015). Using PCA in such cases can lead to results that are unduly influenced by arbitrary data scales and often describe only a few particular outlying points. Thus an active area of research is methods for making PCA robust to such problems (Croux and Ruiz-Gazen, 2005; Higuchi and Jp, 2004; Hubert et al., 2002; Locantore et al., 1999; Maronna, 2005). One approach to dealing with data scale and outliers is data transformation (Zimmerman and Nunez-Anton, 2010; Huang et al., 2008; Hu et al., 2006; Maadooliat et al., 2015). However, choosing the correct transformation is highly application specific and typically entails substantial domain-specific knowledge (Maadooliat et al., 2015). While recommended transformations are established in some fields, determining an appropriate transformation is often itself a substantial question to answer. In this manuscript we tackle this problem by exploring automatic transformation and integration of features as part of a PCA data analysis pipeline. One important area in which automatic transformation of data is necessary is high-content imaging data (Caicedo et al., 2017). Such data consists of hundreds of features extracted from thousands of images using automatic feature detection software. While classically there is only one data matrix to transform, for high-content image data there will be hundreds of matrices to consider: one for each image feature. An appropriate transformation will need to be chosen for each of these feature matrices. In this paper we consider simple and robust ways of adaptively doing this. We show that transforming data can improve visualization and discovery of substantive latent effects in features individually and when integrating together multiple features. Consequently, we explore the interaction of transformation with integrating multiple features to extract a common set of latent variables. As part of all these analyses we employ methods to estimate principal components from data with missing values.

2 Motivating Application: MEMAs

The motivating application of this work is data from the high-content bio-technology called the Microenvironment Microarray (MEMA)(Labarge et al., 2014; Watson et al., 2018; Lin et al., 2017; Smith et al., 2019). MEMAs aid exploration of cellular microenvironment: the cells' immediate physical and bio-chemical surroundings. This microenvironment is important as it is implicated in many cell and tissue level processes, diseases, and dysfunctions (Maman and Witz, 2018; Januschke and Näthke, 2014; Lin et al., 2012; Pelissier et al., 2014; LaBarge, 2013). Better understanding of the microenvironment benefits basic research and furthers an understanding of the interaction between therapeutic agents and regulatory behavior (Teti, 1992; Bissell and Labarge, 2005; LaBarge et al., 2007). Understanding cellular microenvironments has been a long-term research aim of the NIH and consequently is a major component of the LINCS (Library of Integrated Network-Based Cellular Signatures) NIH Common Fund program.

MEMAs facilitate the study of combinations of microenvironmental factors on molecular and biological endpoints via high-throughput image-based profiling of cells. After growing

the cells on the MEMA substrate under various conditions, the cells are stained and imaged with microscopy (Figure 1). The images are then analyzed with software to extract quantitative cellular features. There is one feature matrix for each cellular feature. For example, there will be a feature matrix recording cell size for several hundred MEMA plates (the observations, rows) across several hundred microenvironmental conditions (the variables, columns). The features (and feature matrices) cover a wide range of cellular aspects. There are morphological features like cell area, compactness, eccentricity, perimeter, or solidity. The MEMAs also produce stain intensity features and features capturing the cell cycle state, cell lineage, cell count, texture and many others. Typically, this amounts to several hundred features however the features are not only diverse but may vary from one experiment to the next depending on research interests, software utilized, and underlying biology.

While this plethora of feature data presents new opportunities for discovery, it also necessitates an adaptive approach that can handle a disparate and changing landscape of features. Determining a good transformation for MEMA data is more complicated than traditional -omics experiments due to the number, flexibility, and disparate nature of the features. Consider a nuclei orientation feature (approximately normally distributed) and a DAPI intensity feature (highly right-skewed and always positive, see Supplementary Figures 1 and 2). The appropriate transformation for these two features will likely not be the same since they have very different distributions. We might want to log-transform the intensity feature while leaving orientation feature alone. With this in mind, this paper studies methods for automatically and adaptively choosing appropriate data transformations for any image feature produced by a high-throughput image-analysis platform like a MEMA. Our successful application on MEMAs points to promising applications of this methodology to other image-based cell-profiling technologies like Cyclic Immunofluorescence and Cell Painting (Lin et al., 2016; Bray et al., 2016; Tsujikawa et al., 2017).

3 Methods

In this section we discuss our transformation approach (Section 3.1) and how it helps visualization and discovery of substantive latent effects (Section 3.2). We will also outline calculating the SVD in the presence of missing data (Section 3.3), and how we recover latent effects through data integration (Section 3.4).

3.1 Robust Re-scaling

To process the image-feature matrices we follow three sequential transformation steps:

Procedure 1

Three-step Robust Re-scaling (RR)

-
- Step 1:** (G) robustly “Gaussianize” the data,
Step 2: (Z) convert the data to robust z-scores,
Step 3: (O) remove outliers.
-

These three steps are applied to each of the feature matrices individually.

3.1.1 The Gaussianizing Step (G)—The (G), or “Gaussianizing”, step transforms the data using a robust Box-Cox-like procedure (Box and Cox, 1964). The traditional Box-Cox procedure for finding an optimal transformation is not well suited for this data because it is not robust. Indeed, central to the Box-Cox estimator is a re-scaled sample variance. As the breakdown point of the sample variance is exactly zero, we expect the breakdown point of the Box-Cox estimator to also be near zero (Marazzi and Yohai, 2004; He et al., 1990; Cook and Wang, 1983; Atkinson, 1986). A breakdown point near zero indicates that even one outlier can lead to a drastically over-fit Box-Cox transformation. To overcome this, the (G) step uses an interpretable and robust “divide-and-summarize” approach to improve stability and avoid over-fitting. The procedure first divides the data column-wise and estimates a Gaussianizing (Box-Cox) transformation for each column of the feature matrix. To avoid over-fitting, the procedure then robustly summarizes the transformation parameters, choosing the median among the column-wise estimates. This median transformation is then used to transform the original feature matrix. This model-averaging approach is similar to attribute bagging seen in supervised learning applications (Bryll et al., 2003). However because all attributes (columns) are used in (G), as opposed to a random subspace procedure, the estimates we calculate are interpretable. If we consider a model like that of Carroll (1980) where a certain percentage of the data is arbitrarily corrupted, then by using the consensus median estimate of the transformation parameter, up to 50% of the columns of the data can be arbitrarily corrupted without arbitrarily corrupting the estimate of the transformation.

Consider MEMA data where each column is a different microenvironment being studied. Our final transformation is a transformation that works well for the typical (median) microenvironment and ignores any aberrant microenvironments. This allows (G) to avoid undue influence of technical problems. We will see in Section 4 that this avoids technical problems of the NID and ELN microenvironments. Furthermore, the extreme transformations proposed by these two microenvironments allows us to flag them for further quality-control scrutiny.

Let $Y \in \mathbb{R}^{M \times N}$ be a specific feature matrix and let $\mathcal{T}_1, \dots, \mathcal{T}_Q$ be a collection of Q parameterized transformation families so that for any $q = 1, \dots, Q$ the family $\mathcal{T}_q = \{T_\lambda^{(q)} \mid \lambda \in \Lambda^{(q)}\}$ consists of differentiable, monotonic, transformations $T_\lambda^{(q)}: S \rightarrow \mathbb{R}$ on some $S \subseteq \mathbb{R}$. The goal is to optimize over the union of these families and choose the transformation that makes the data close to being normal (without over-fitting).

In our application (Section 4) we will choose $Q = 2$ families over which to search: a power family $\mathcal{T}_1 = \mathcal{T}_{power} = \{(\text{sign}(y)|y|^\lambda - 1)/\lambda, \lambda \in \mathbb{R}\}$ and an arc-hyperbolic-sine family $\mathcal{T}_2 = \mathcal{T}_{ahs} = \{\sinh(\lambda y)/\lambda, \lambda \geq 0\}$. We choose these two families because they cover a range of power and sigmoidal shapes. While the family of power transformation is versatile, it cannot deal elegantly with negative values in the data. Thus, we have included the archyperbolic-sine family because it is a well-studied family of transformations that can

deal with negative values (Nowicka et al., 2019). Many reasonable choices of parameterized families can be made and nothing in our discussion depends on the specific choices. We include other options of families in our software.

Before we describe the procedure for optimizing over many families, we will first consider the simpler case when $Q = 1$ and discuss how to choose an optimal transformation over a single family generically denoted $\mathcal{T} = \{T_\lambda \mid \lambda \in \Lambda\}$. Define Y_{*j} be the j^{th} column of Y , and for any $\lambda \in \Lambda$ let $Y_{*j}(\lambda) = T_\lambda(Y_{*j})$ be Y_{*j} under the transform T_λ . The goal is to choose a $\hat{\lambda}$ so that $Y_{*j}(\hat{\lambda})$ is approximately normally distributed for each $j = 1, \dots, N$. The (G) approach follows two steps:

For the first step, estimate $\hat{\lambda}_j$ using the traditional Box-Cox approach on Y_{*j} . Assume

Procedure 2

(G) Transformation Estimation for One Family ($Q = 1$)

Step 1: (divide, estimate) determine the optimal $\hat{\lambda}_j$ for each column Y_{*j} so that $Y_{*j}(\hat{\lambda}_j)$ is as normal as possible,

Step 2: (summarize) set $\hat{\lambda} = \text{median}_j \hat{\lambda}_j$

there is some λ_j so that $Y_{ij}(\lambda_j) \stackrel{iid}{\sim} N(\mu_j, \sigma_j^2)$ for $\mu_j \in \mathbb{R}$ and $\sigma_j \geq 0$. Then let $\hat{\lambda}_j$ be MLE of λ_j . This is obtained by profiling the likelihood over μ_j and σ_j^2 and then maximizing the profile likelihood over λ_j . If L_j is the profile likelihood of λ_j profiling over μ_j and σ_j^2 then $\hat{\lambda}_j \stackrel{def}{=} \arg\max_{\lambda_j \in \Lambda} L_j(\lambda_j)$.

After estimating $\hat{\lambda}_j$ for each column, the second step is to summarize the collection of $\hat{\lambda}_j$'s into a single $\hat{\lambda}$. This is done with the median. Define $\hat{\lambda}$ as the element-wise median $\hat{\lambda} = \text{median}_j \hat{\lambda}_j$.

Concisely, the procedure when $Q = 1$ is to first divide and optimize within each column and then median-summarize across the column-wise estimates. When $Q > 1$ we add an additional step to first determine which family among $\mathcal{T}_1, \dots, \mathcal{T}_Q$ is best. The procedure is described in Procedure 3.

Procedure 3

(G) Trans. Estimation for Multiple Families ($Q > 1$)

Step 1: determine which family is the best over-all, call it $\hat{q}$

Step 2: estimate λ using just the optimal family $\mathcal{T}_{\hat{q}}$ (following the previous procedure of Procedure 2 for $Q = 1$).

This procedure first determines the best family individually for each column and then uses the family that is best among a plurality of columns. More specifically, let $L_j(q_j, \lambda_j)$ be the likelihood of the j^{th} column after transformation using the q_j^{th} family

and transformation parameter λ_j . Optimize $L_j(q_j, \lambda_j)$ jointly over λ_j and q_j and let $(\hat{q}_j, \hat{\lambda}_j) = \arg \max_{\lambda_j \in \Lambda^{(q)}} L_j(q_j, \lambda_j)$ and $\hat{q} = \text{mode}_j \hat{q}_j$ so that $\hat{q}$ is the family that is the best among a plurality of the columns. Once we have determined this optimal family $\hat{q}$ we then estimate $\hat{\lambda}$ following the procedure when $Q = 1$ using the family $\mathcal{T}_{\hat{q}}$. Finally, define the Gaussianized version of Y as $G(Y) \stackrel{\text{def}}{=} T_{\hat{\lambda}}^{(\hat{q})}(Y)$.

3.1.2 The z-score Step (Z)—The second step in the (RR) procedure is the (Z) step, a robust z-score transformation. Let $\tilde{Y}$ be a vectorized version of Y and $\tilde{Y}_{(q)}$ be the q -winsorized version of $\tilde{Y}$. In this paper we will use $q = 0.001$ replacing everything below the q^{th} quantile of $\tilde{Y}$ by the q^{th} quantile and replacing everything above the $(1 - q)^{\text{th}}$ quantile of $\tilde{Y}$ by the $(1 - q)^{\text{th}}$ quantile. Given this, the robust z-score version of Y is defined as $Z(Y) \stackrel{\text{def}}{=} (Y - \hat{\mu})/\hat{\sigma}$ where $\hat{\mu}$ and $\hat{\sigma}$ are mean and s.d. estimates of $\tilde{Y}_{(q)}$. Notice that the final values $Z(Y)$ have not been winsorized themselves, the winsorization has only been used in the calculation of $\hat{\mu}$ and $\hat{\sigma}$.

3.1.3 The Outlier Removal Step (O)—The final of the three (RR) steps is outlier removal. The outlier removal procedure simply thresholds z-scores and marks as missing anything beyond four standard deviations. First let $Z(Y)$ be the robust z-scored version of Y . We then define $O(Y)_{ij} = Y_{ij}$ if $|Z(Y)_{ij}| \leq z$, and $O(Y)_{ij} = \text{“NA”}$, otherwise. Here, “NA” denotes a missing value. To be conservative in this paper we use $z = 4$ although this is somewhat arbitrary. With $z = 4$ if the data is truly normal this removes only about 3e-3 percent of the data from each tail.

3.1.4 The Three-Step (RR) Procedure—Given these definitions, the three step (RR) transformation is to apply the (G), (Z), and then (O) transformations. If Y is a feature matrix then we define $RR(Y)$ as $RR(Y) = O(Z(G(Y)))$.

3.2 Transformations and Latent Effects

A central component in data analysis is the identification of important latent effects both visually and quantitatively. For MEMAs, we divide latent effects into two categories: (1) biological effects and (2) technical effects. Biological effects include, for example, differences in biological endpoints due to ECMps or ligands. Technical effects are unwanted and we are interested in identifying them so that we may remove them. Examples include batch across plates or spatial effects within wells. Discovery of latent effects is typically done through visual inspection of plots or quantitative analysis like PCA. Unfortunately, methods like PCA are often misled by prominent aspects of the data are unrelated to substantive latent effects. As an example, consider how PCA can be misled when used to identify groups in skewed data. Let $u^{(1)}, u^{(2)}, v^{(1)}, v^{(2)} \in \mathbb{R}^N$ have elements that are *i.i.d* from a standard log-normal distribution. For a small $\delta \in \mathbb{R}$ and noise $\epsilon \in \mathbb{R}^{2N \times N}$ define a block data matrix Y as $Y = \begin{bmatrix} A \\ B \end{bmatrix} + \epsilon$ where $A = u^{(1)}v^{(1)'} + \delta u^{(2)}v^{(2)'}$ and $B = \delta + u^{(2)}v^{(2)'}$ so that the first N rows of the data matrix and the last N rows of the data matrix constitute two groups

with a mean difference of δ . The left side of Figure 2 displays a histogram of the elements of Y for a simulation using $\delta = 1/2$ and distributed *i.i.d.* standard normal. Visually, it is difficult to distinguish between the two groups in the left-hand panel of Figure 2 because the group difference is over-shadowed by the data's long tails. Consequently, PCA identifies the variance due to tail skewness, not the group difference, as the most prominent variation in the data. While the first two principal components capture more than 99% of the total variance in this example, they only capture about 50% of the group difference (Suppl. Figure 3). The right side of Figure 2 shows that a log transformation makes the groups more prominent through visual inspection. The transformation also helps quantitative analyses. The transformation un-skews the data thereby attenuating the effect of the tails on PCA. In this case, while the first two PCs only capture about 80% of the total variation, they capture about 94% of the variation due to group difference (Suppl. Figure 3).

Motivated by the previous example, we want to attenuate the influence of prominent, yet uninformative, variation when visualizing data and when applying quantitative methods. The (RR) transformation does this by ameliorating the effects of two commonly encountered, and potentially misleading, aspects of data: (1) skewness in measurement scales and (2) anomalous outliers. By anomalous outliers we mean extremely unusual data points that are not informative of much beyond their own uniqueness. For example, cells may have difficulty growing on a spot, or software might produce image-analysis artifacts like segmentation anomalies.

To guard against non-substantive variation (RR) applies three robust re-scaling steps. (G) robustly prevents a feature's naturally long-tailed measurement scale from dominating analysis by de-skewing each feature's distribution. While the traditional Box-Cox procedure is likely to over-fit in the presence of genuine outliers, (G) will not. Indeed, traditional Box-Cox will propose an extreme transformation to rectify even a single outlier. (G) only transforms the data to un-skew fundamentally skewed data, not simply to reign in a few points. Instead, outliers are handled more parsimoniously by a procedure that specifically targets them. To remove outliers (RR) first converts the data to robust z -scores using (Z) and then removes any entry of the feature matrix bigger in magnitude than four with (O).

3.3 Complete Singular Vectors

A common statistical tool used to recover latent effects is the singular value decomposition (SVD). For a data matrix Y with a singular value decomposition (SVD) of $Y = U\Sigma V'$ we call the columns of the U left singular vectors and the columns of V the right singular vectors. If one were to mean-center the columns of Y , these singular vectors define the principal components (PCs). We avoid this term because we do not mean-center.

When calculating the SVD for image-based data we often need to account for missing data. Missing values arise for domain-specific reasons (e.g. cells failed to grow on a MEMA), image-analysis reasons (e.g. the software could not detect any features), and because (RR) introduces missing values as part of (O). As the SVD is undefined for matrices with missing values, we use "complete" singular vectors calculated from re-scaled pairwise-complete gram matrices. This is similar to pairwise-complete covariance matrices (e.g. `cor` in R).

Let $Y \in \mathbb{R}^{M \times N}$ be a feature matrix with missing values. Define Y_0 as Y with missing values replaced by zeros and Y_1 so that $(Y_1)_{ij} = \mathbb{1}\{Y_{ij} \text{ is not missing}\}$. We define the re-scaled pairwise-complete left Gram matrix $Y \cdot Y'$ so that

$$(Y \cdot Y')_{ij} = \frac{N}{n_{ij}} \sum_{k=1}^N (Y_0)_{ik}(Y_0)_{jk}$$

where $n_{ij} = (Y_1 Y_1')_{ij}$ is the number of pairwise-complete entries between row i and j of Y .

$Y \cdot Y'$ is a matrix of re-scaled inner products of the rows of Y accounting for the number of non-missing pairs between rows. We similarly define the right gram matrix $Y' \cdot Y$ replacing Y for Y' above. We call the eigenvectors of $Y \cdot Y'$ and $Y' \cdot Y$ the complete left/right singular vectors. WLOG they are ordered decreasing by eigenvalue and we keep only those with positive eigenvalues. For brevity we henceforth omit the adjective “complete”, referring to these simply as the “singular vectors.” If there are no missing values they are identical. While other methods exist for calculating the SVD on matrices with missing values, (e.g. Hastie et al. (2015)), because the total number of missing values is small for MEMA data, our simple procedure will allow quick and intuitive calculations of singular vectors that are highly commensurate existing approaches.

3.4 Average Singular Vectors

In addition to recovering important latent effects in individual features, we are interested in latent effects common to multiple features. To extract a common set of latent effects from a collection of P features $Y^{(1)}, \dots, Y^{(P)}$ we use the eigenvectors from the average of their re-scaled pairwise-complete left and right gram matrices, respectively,

$$\frac{1}{P} \sum_{p=1}^P Y^{(p)} \cdot Y^{(p)'} \text{ and } \frac{1}{P} \sum_{p=1}^P Y^{(p)'} \cdot Y^{(p)}.$$

We call these eigenvectors the left and right average singular vectors (ASVs).

4 Application to MEMA Data

4.1 Structure of MEMA Data

We work with MEMA data from the MEP-LINCS Center at the Oregon Health and Science University. The data is accessible through Synapse with identifiers syn10155286, syn10155289 and syn10155292 (Gray et al., 2014) available at http://www.synapse.org/MEP_LINCS. In total we analyze 24 MEMAs of human epithelial mammary tissue (MCF10A). The 24 MEMAs come in three batches of eight plates. Each MEMA plate is divided evenly into eight wells. Each well contains 700 spots in a 20 by 35 grid. Cells are added to the wells and bind to the spots. Subsequently, a buffer solution containing a specific ligand is added to each well. Thus the cells can grow out in the presence of different ECM proteins and ligands. The pattern of ECMps is identical across all wells (see Supplementary Figure 4) however a (potentially) different ligand is added to each well (see Supplementary

Figure 5). After incubating the cells for 72 hours they are fluorescently stained, imaged, and cell-level features are extracted with image analysis software. For the analysis in this paper, we work with spot-level features (median summarized cell-level features). For each image feature we have a data matrix of 192 wells (3 batches $\times$ 8 plates $\times$ 8 wells) by 694 spots (we remove 6 alignment spots with no cells from the 700). In total we will work with 103 image features (Suppl. Table 1).

4.2 Features and Transformations Considered

The MEMA plates we analyze are grown, stained and imaged in three separate processing batches. A different set of stains is used in each batch. Those sets are: (1) “SS1” (containing stains DAPI, Actin, CellMask and MitoTracker), (2) “SS2noH3” (containing stains DAPI, Fibrillarin and EdU), and (3) “SS3” (containing stains DAPI, KRT5, KRT19 and CellMask). Because each of these batches use a different staining set we refer to them as the “staining batches.” While these batches are separate experiments, aside from the staining set the experimental conditions were made as identical as possible.

In total there are 103 different image features extracted from the MEMAs. A different set of features is extracted in each staining batch with some being common across multiple batches. There are 50 features extracted in at least two of the staining batches and 18 features that are extracted from all three. We focus on four features in this paper: (1) cell area (notated on synapse as “Cells CP AreaShape Area”) (2) cell compactness (“Cells CP AreaShape Compactness”) (3) spot cell count (“Spot PA SpotCellCount”) (4) total cytoplasm DAPI intensity (“Cytoplasm CP Intensity IntegratedIntensity Dapi”). We choose these features because they represent several different feature types. The first two are morphological traits of cells, the third is the cell count, and the last is an intensity. We have deposited the full results for all features at [gjhunt.github.io/rr](https://github.com/gjhunt/rr). To explore the effects of (G), (Z) and (O), we consider five transformations of the features: (1) no transformation (NT), (2) the (G) step only (3) the (Z) step only (4) the (O) step only (5) the three-step (RR) transformation.

4.3 Visualization

4.3.1 Feature Distributions—A typical first step in exploratory analysis is data visualization. Simple data visualizations can succinctly summarize the major features of the data and inform qualitative analyses. In Figure 3 we plot the distribution of cell area for the five transformations. The densities correspond to staining batches. The bold line is the density of all data combined. Notice in Figure 3 that the density of (NT) largely reflects the data’s long tail. The same can be said for (Z). Conversely, the other transformations reveal the staining batches. Both (O) and (G) de-emphasize the data’s long tail in favor of the group difference. Furthermore, in (RR) these groups are approximately Gaussian. Suppl. Figures 6-9 shows a similar plots for other features and effects.

4.3.2 Heat-maps—Another way to visualize the MEMA data is through heat-map pseudo-images. These pseudo-images are heat-maps of the value of a feature for each spot plotted following the same physical layout as the MEMA. These pseudo-images can be useful for discovering spatial effects and assessing the quality of data. As an example, we

visualize cell area this way in Supplementary Figure 10. In Figure 4 (A) we display a single well from Supplementary Figure 10 across the five transformations. The colors are more blue if they are close to the minimum cell area, red if they are close to the maximum, and white if they are half-way between. Dark grey spots are omitted according to the MEMA design. Note that the color scale is determined globally over all wells and all plates in the dataset (Suppl. Figure 10).

This figure is not very informative for (NT) or (Z). The skewness and outliers assign the bulk of data to a tiny range of colors meaning the plots are essentially a single color. Conversely, for (G) and (O) we see a spatial effect between the right and edges and the rest of the well. We also see a non-spatial effect where certain spots are much different than their surroundings. We circle these spots in orange in Suppl. Figure 11. In Section 4.6 we show that this is an effect of the ECMps NID1 and ELN. We can see from these plots that (RR) strongly highlights the spatial effects as well as the NID/ELN effect.

In Figure 4 (B) we focus on a different well of Suppl. 10. Here, the green spots indicate missing data. These spots are missing either due to experimental error or because they have been removed as part of analysis. We see similar behavior where (G) and (O) reveal spatial differences between the upper right and the rest of the well. This spatial effect is also seen in (RR) however the number of points removed is much different in (RR) compared to (O). This highlights the difference between (O) thresholding outliers without transformation and (RR) thresholding outliers after (G). We believe thresholding based on a z -score makes most sense on a Gaussianized scale (as in (RR)). Note also that outliers are defined in a global context of the entire data, so while many values are marked as outliers by (RR) in this particular well it is a small percentage of the entire data.

In addition to highlighting spatial effects, these transformations also reveal batch effects between plates, wells, and staining batches. In Figure 5 we display the heat-map pseudo-image of cell area for eight wells across (NT), (G), (O) and (RR). ((Z) is identical to (NT).) The top four wells in each sub-plot are from the first staining batch, the bottom four wells are from the second. Nonetheless, we see little indication of batch in (NT). However batch is visible in (G), (O) and (RR). The bottom of (G) is lighter blue than the top, and the top of (O) is lighter red than the bottom. In (RR) we have solid-blue in the top and mostly red in the bottom. Being better able to identify batch effects hopefully will aid down-stream procedures to account for such effects.

4.4 Recovering Technical Effects Across Wells

Batch effects are a common and well-studied problem in high-throughput biological experiments like MEMAs (Leek et al., 2010). Often, such batch effects obscure biological variation of interest. To deal with this problem, unwanted variation, like batch, is typically identified using the SVD and projected out of the data. In this section we explore how (RR) helps identify unwanted variation like batch using the SVD. We focus on the large staining batch effect as it was visible by eye in Figure 5.

We assess the transformations by measuring the percentage of the batch captured by the first k singular vectors of the transformed feature matrix. Let $U = [u_1, \dots, u_N] \in \mathbb{R}^{M \times N}$ be

the (complete) left singular vectors of a feature matrix and $B \in \mathbb{R}^{M \times D}$ be the batch (A) indicator matrix so that $B_{ij} = 1$ if well i is in batch j for $j = 1, \dots, D$. Here we have $D = 3$ for the three staining batches. For $k = 1, \dots, N$ and $t = 1 \dots \min(k, D)$ define $C_k^{(t)}$ to be the t^{th} canonical correlation between the first k left singular vectors $U_k = [u_1, \dots, u_k]$ and the batch B . Then let $C_k^2 = \frac{1}{D} \sum_{t=1}^{\min(k, D)} (C_k^{(t)})^2$ to be the average of these squared canonical correlations. We can interpret C_k^2 as the percentage of the batch B that is captured by these first k singular vectors. In Figure 6 (A) we plot C_k^2 on the y -axis and vary k across the x -axis from $k = 1$ to 192. From this figure we see that the transformations enhance identification of the staining batch. Consider the cell area and total DAPI intensity features. As compared with no transformation (NT), these plots shows that (G), (O), and (RR) increase how much of the staining batch is captured by the first several singular vectors. These transformations attenuate the non-informative tails of the distributions and focus the singular vectors on the differences across the staining batches.

We summarize batch recovery for all features in Figure 6 (B). Here, we calculate the area under the CC curves (AUC) for each feature as $AUC = \sum_{k=1}^{192} C_k^2$. Broadly, we see the same behavior in Figure 6 (B) as displayed in Figure 6 (A). (RR) seems to generally improve recovery of the staining batch. Sometimes we see a substantial improvement (e.g. Cell Compactness) and rarely do we see that (RR) is detrimental. In Suppl. Figures 15 - 20 we display similar plots for the recovery of plate, well, and ligand effects.

4.5 Data Integration for Discovering Between-Well Effects

Given the close relationship among many of the MEMA image features, latent effects that appear in one feature may show up in other features. Shared effects give insight into biological and technical aspects that are important across many features. To extract these common effects we integrate information across MEMA features using the left average singular vectors (ASVs) as described in Section 2.

In the left panel of Figure 7 we plot the mean squared canonical correlations (C_k^2) between the first k left ASVs and the staining batch. We calculate the ASVs using the 18 features measured in every plate. From this figure we can see that (Z) and (RR) quickly and strongly recover the staining batch. The AUC for these curves is in excess of 0.95, meaning it recovers batch better than the majority of individual features. The ASVs “average-out” feature-specific effects and amplify common effects like staining batch.

It is notable that the (Z) and (RR) recover batch significantly better than (O), (G), and (NT). This happens because the ASVs element-wise average Gram matrices across features. If these Gram matrices are on vastly different scales their average is biased towards the largest features. This arbitrarily weights features’ by their scales. To equitably integrate information all features should have values in a similar range as in (Z) and (RR). Thus they recover batch better.

Finally, it is notable how well (Z) does alone. This happens because the averaging used to compute the ASVs conveys some of the same benefits as (G) and (O). This is true so long as we do not have a small number of features or systematic skewness or outliers across features. In the right panel of Figure 7 we calculate the ASVs using only five features with several high-leverage points. Here we see a separation between (Z) and (RR) since the average is over a small number of highly-skewed features. In any case, including (G) and (O) steps does not seem to hurt the analysis and thus we still recommend the full three-step (RR) transformation for integrating features in this manner. Similar, but attenuated results for plate, well and ligand are shown in Supplementary Figures 21 - 23.

4.6 Discovering Biological and Spatial Effects within Wells

The left singular vectors of the feature matrices reveal latent effects across the wells, plates, and staining batches. Similarly, the right singular vectors (RSVs) reveal effects across the spots. In Figure 8 (A) we display a scatter plot of the first four RSVs of total cytoplasmic DAPI intensity for (NT) and (RR).

A prominent feature of Figure 8 (A) is the separation between the ECMps ELN, NID1, and the rest. Upon further investigation of the underlying MEMA images we find that this effect manifests because the cells have difficulty adhering to the ELN and NID1 ECMp substrates. Notice the cell count heat-map in Suppl. Figure 13 shows that the cell count in the ELN and NID1 spots are significantly lower than other spots. While this ELN-NID1 effect is present in the un-transformed data, it is more prominent in (RR). The first RSV from the un-transformed data does capture the effect, however the second through fourth RSVs are focused on explaining several outliers. Moreover, (RR) separates NID1 and ELN from the other ECMps and from each other.

In Figure 8 (B) we plot pseudo-image heat-maps of the first ten RSVs for cytoplasmic DAPI intensity arranging elements of the RSVs according to the MEMA plate spatial layout. In addition to the ELN/NID effect, these plots reveal common spatial patterns across wells. These patterns are more visible for (RR) than (NT) as the RSVs of (NT) mostly capture outliers. It is important to identify such unwanted effects so that we can properly account for them downstream. In Supplementary Figures 25 - 29 we display similar scatter plots and heat-maps for the other example features. They tell similar stories.

To see what biological effects can be found if we *a priori* remove the dominating ELN-NID effect, we re-analyze the MEMA data after removing these spots. Now we find an effect separating THBS from the other ECMps. This is particularly prominent in morphological features. As an example, in Figure 9 we plot a scatter plot of the top four RSVs for cell compactness. (RR) reveals a difference between THBS and the other ECMps. This effect shows up in many of the morphological features but not cell count. Thus this THBS effect does not appear to be of similar origin to the ELN-NID effect. Instead, it appears to be a biological effect on cell morphology.

4.7 Data Integration for Discovering Within-Well Effects

In section 4.5 we saw that data integration helped make salient important between-well effects. In a similar fashion, the average right singular vectors (ASVs) help bring out within-

well effects. In Figure 10 (A) we plot the first two right ASVs against each other. Again, (RR) equitably integrates information from all features and helps highlight the NID/ELN effect. Finally, we display heat-maps pseudo-images for the first ten right ASVs in Figure 10 (B). The right ASVs for (NT) seem to be mostly picking up a couple of outliers. On the other hand, (RR) strongly picks up several interesting spatial effects.

5 Discussion

In this paper we have explored the effects of several transformations as part of a pre-processing pipeline of high-content image data, in particular, data from microenvironment microarrays. The goal of these transformations is to emphasize important latent effects in the data and attenuate common and misleading aspects.

Un-transformed feature data is often encumbered by skewed measurement scales, outliers, or both. These aspects can hinder discovery of substantive latent effects. To de-emphasize such misleading aspects of the data (O), (G), and their combination in (RR) were helpful. (O) removed outliers using a conservative threshold and (G) reduced skewness by Gaussianizing the data. Additionally, (RR) included a (Z) step that converted values to robust z -scores. (Z) and (RR) allowed features to be straight-forwardly integrated with a simple arithmetic average of Gram matrices. In the analysis of MEMA data we showed that a combination of a Gaussianizing transformation (G), z -score transformation (Z), and removal of outliers (O) can improve visualization and discovery of biological and technical latent effects in both features individually and when combining features together. Finally, as (RR) automatically chose transformations for each feature this allowed adaptive application of (RR) to a data containing many different features. This adaptive ability makes (RR) a promising candidate for data generated by other image-based technologies like Cyclic Immunofluorescence (CycIF) or Cell Painting (Lin et al., 2016; Bray et al., 2016; Tsujikawa et al., 2017). For example, CycIF, through a series of imaging and washing steps, allows up to 30-channel immunofluorescent imaging and thus extraction of potentially several hundred features. Like MEMAs, each of these features will need a transformation to be adaptively chosen. Exploring the application of (RR) to other high-content and highly-multiplexed technologies is a direction we hope to explore in future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

The authors gratefully acknowledge support from the National Science Foundation (grant no. DMS-1646108) and the National Institutes of Health (grants nos. U54HG008100 and 1U54CA209988).

References

- Atkinson AC (1986). Diagnostic Tests for Transformations. *Technometrics*, 28(1), 29–37.
- Bissell MJ and Labarge MA (2005). Context, tissue plasticity, and cancer: are tumor stem cells also regulated by the microenvironment? *Cancer cell*, 7(1), 17–23. [PubMed: 15652746]

- Box GEP and Cox DR (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2), 211–252.
- Bray M-A, Singh S, Han H, Davis CT, Borgeson B, Hartland C, Kost-Alimova M, Gustafsdottir SM, Gibson CC, and Carpenter AE (2016). Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9), 1757–1774. [PubMed: 27560178]
- Bryll R, Gutierrez-Osuna R, and Quek F. (2003). Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognition*, 36(6), 1291–1302.
- Caicedo JC, Cooper S, Heigwer F, Warchal S, Qiu P, Molnar C, Vasilevich AS, Barry JD, Bansal HS, Kraus O, Wawer M, Paavolainen L, Herrmann MD, Rohban M, Hung J, Hennig H, Concannon J, Smith I, Clemons PA, Singh S, Rees P, Horvath P, Linington RG, and Carpenter AE (2017). Data-analysis strategies for image-based cell profiling. *Nature Methods*, 14(9), 849–863. [PubMed: 28858338]
- Carroll R. (1980). A robust method for testing transformations to achieve approximate normality. *Journal of the Royal Statistical Society: Series B*, 42(1), 71–78.
- Cook RD and Wang PC (1983). Transformations and Influential Cases in Regression. *Technometrics*, 25(4), 337–343.
- Croux C. and Ruiz-Gazen A. (2005). High breakdown estimators for principal components: The projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1), 206–226.
- Gray J, Heiser L, and Korkola J. (2014). Microenvironment Perturbagen (MEP) LINC on Synapse.
- Hastie T, Mazumder R, Lee JD, and Zadeh R. (2015). Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. *Journal of Machine Learning Research*, 16, 3367–3402. [PubMed: 31130828]
- He X, Simpson DG, and Portnoy SL (1990). Breakdown Robustness of Tests. *Journal of the American Statistical Association*, 85(410), 446.
- Higuchi I. and Jp EA (2004). Robust Principal Component Analysis with Adaptive Selection for Tuning Parameters Shinto Eguchi. Technical report.
- Hu J, Wright FA, and Zou F. (2006). Estimation of expression indexes for oligonucleotide arrays using the singular value decomposition. *Journal of the American Statistical Association*, 101(473), 41–50.
- Huang JZ, Shen H, and Buja A. (2008). Functional principal components analysis via penalized rank one approximation. *Electronic Journal of Statistics*, 2, 678–695.
- Hubert M, Rousseeuw PJ, and Verboven S. (2002). A fast method for robust principal components with applications to chemometrics. In *Chemometrics and Intelligent Laboratory Systems*, volume 60, pages 101–111.
- Hubert M, Rousseeuw P, and Verdonck T. (2009). Robust PCA for skewed data and its outlier map. Technical report.
- Januschke J. and Näthke I. (2014). Stem cell decisions: A twist of fate or a niche market? *Seminars in Cell & Developmental Biology*, 34, 116–123. [PubMed: 24613913]
- LaBarge M. (2013). Breaking the Canon: Indirect Regulation of Wnt Signaling in Mammary Stem Cells by MMP3. *Cell Stem Cell*, 13(3), 259–260. [PubMed: 24012364]
- LaBarge MA, Petersen OW, and Bissell MJ (2007). Of microenvironments and mammary stem cells. *Stem cell reviews*, 3(2), 137–46. [PubMed: 17873346]
- Labarge MA, Parvin B, and Lorens JB (2014). Molecular deconstruction, detection, and computational prediction of microenvironment-modulated cellular responses to cancer therapeutics. *Advanced drug delivery reviews*, 69–70, 123–31.
- Leek JT, Scharpf RB, Brannan HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, and Irizarry RA (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10), 733–739.
- Lin C-H, Lee JK, and LaBarge MA (2012). Fabrication and Use of MicroEnvironment microArrays (MEArrays). *Journal of Visualized Experiments*, (68), 1–7.
- Lin C-H, Jokela T, Gray J, and LaBarge MA (2017). Combinatorial Microenvironments Impose a Continuum of Cellular Responses to a Single Pathway-Targeted Anti-cancer Compound. *Cell Reports*, 21(2), 533–545. [PubMed: 29020637]

- Lin J-R, Fallahi-Sichani M, Chen J-Y, and Sorger PK (2016). Cyclic Immunofluorescence (CycIF), A Highly Multiplexed Method for Single-cell Imaging. In *Current Protocols in Chemical Biology*, volume 8, pages 251–264. John Wiley & Sons, Inc., Hoboken, NJ, USA. [PubMed: 27925668]
- Locantore N, Marron JS, Simpson DG, Tripoli N, Zhang JT, Cohen KL, Boente G, Fraiman R, Brumback B, Croux C, Fan J, Kneip A, Marden JI, Peña D, Prieto J, Ramsay JO, Valderrama MJ, Aguilera AM, Locantore N, Marron JS, Simpson DG, Tripoli N, Zhang JT, and Cohen KL (1999). Robust principal component analysis for functional data. *Test*, 8(1), 1–73.
- Maadooliat M, Huang JZ, and Hu J. (2015). Integrating Data Transformation in Principal Components Analysis. *Journal of Computational and Graphical Statistics*, 24(1), 84–103. [PubMed: 25914514]
- Maman S. and Witz IP (2018). A history of exploring cancer in context. *Nature Reviews Cancer*, 18(6), 359–376. [PubMed: 29700396]
- Marazzi A. and Yohai VJ (2004). Robust Box-Cox Transformations for Simple Regression. In *Theory and Applications of Recent Robust Methods*, pages 173–182. Birkhäuser Basel, Basel.
- Maronna R. (2005). Principal components and orthogonal regression based on robust scales. *Technometrics*, 47(3), 264–273.
- Nowicka M, Krieg C, Crowell HL, Weber LM, Hartmann FJ, Guglietta S, Becher B, Levesque MP, and Robinson MD (2019). CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Research*, 6, 748.
- Pelissier FA, Garbe JC, Ananthanarayanan B, Miyano M, Lin CH, Jokela T, Kumar S, Stampfer MR, Lorens JB, and LaBarge MA (2014). Age-Related Dysfunction in Mechanotransduction Impairs Differentiation of Human Mammary Epithelial Progenitors. *Cell Reports*, 7(6), 1926–1939. [PubMed: 24910432]
- Smith R, Devlin K, Kilburn D, Gross S, Sudar D, Bucher E, Nederlof M, Dane M, Gray JW, Heiser L, and Korkola JE (2019). Using Microarrays to Interrogate Microenvironmental Impact on Cellular Phenotypes in Cancer. *Journal of visualized experiments : JoVE*.
- Teti A. (1992). Regulation of Cellular Functions by Extracellular Matrix. Technical report.
- Tsujikawa T, Kumar S, Borkar RN, Azimi V, Thibault G, Chang YH, Balter A, Kawashima R, Choe G, Sauer D, El Rassi E, Clayburgh DR, Kulesz-Martin MF, Lutz ER, Zheng L, Jaffee EM, Leyshock P, Margolin AA, Mori M, Gray JW, Flint PW, and Coussens LM (2017). Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated with Poor Prognosis. *Cell reports*, 19(1), 203–217. [PubMed: 28380359]
- Watson SS, Dane M, Chin K, Jonas O, Gray JW, and Korkola JE (2018). Microenvironment-Mediated Mechanisms of Resistance to HER2 Inhibitors Differ between HER2+ Breast Cancer Subtypes. *Cell Systems*, 6, 329–342.e6. [PubMed: 29550255]
- Zimmerman DL and Nunez-Anton VA (2010). Antedependence models for longitudinal data. Chapman & Hall/CRC.

Figure 1:

MEMAs: a plastic substrate partitioned into wells. Each well is an array of hundreds of $\sim 400\mu\text{m}$ printed “spots”. Added cells randomly bind to spots and interact with: (1) an extra-cellular matrix protein (ECMp) specific to the spot and (2) a ligand specific to a well. After growing, cells are immunofluorescently stained, imaged, and quantified cellular features extracted.

Figure 2: Skewed data (Left) before a transformation and (Right) after a log-transformation. The transformation makes the group difference more prominent.

Cells_CP_AreaShape_Area

Figure 3:

Density of elements of cell area feature matrix. Bold density is all elements combined. Other densities are the densities for the two staining batches. Subplots are for five processing transformations.

Figure 4:

Heat map of cell area for two wells (A) and (B) across the five transformations (NT), (G), (Z), (O), (RR). Sub-plots of Supplementary Figure 10. Color scaled is determined globally over all spots, wells, and plates in the dataset to reflect the fact that the transformation is similarly calculated over this data. Thus we see no blue in this (NT) sub-plot as we see almost no blue in Supplementary Figure 10. This plot is a representative microcosm of the larger plot. Green indicates missing data.

Figure 5:
Heat map of eight wells across the five transformations (NT), (G), (Z), (O), (RR). Top row of each subplot is from first staining batch. Bottom row is from second staining batch. Colors are more blue if they are close to the minimum, red if they are close to the maximum, and white if they are close to half-way between. Green spots are missing. Dark grey spots are omitted according to the MEMA design.

Figure 6:

(A) Mean of the squared canonical correlations between the first k left singular vectors and the staining batch dummy variables. (B) Grand mean of the squared canonical correlations across number of components. Features ordered left to right decreasing by the difference in the AUC between (RR) and (NT). Thus those on the left (RR) performs relatively better than (NT).

Figure 7:

Mean of the squared canonical correlations between the first k average left singular vectors and the staining batch dummy variables. The average left singular vectors come from integration of (Left) the 18 features that are measured across all MEMAs, and, (Right) the five with the highest leverage points (among those 18).

Figure 8:

(A) Scatter plot of elements of first four right singular vectors against each other for the total cytoplasmic DAPI intensity feature. Shape and color indicate ECMp of the spot corresponding to the elements of the singular vector. (B) Heat map of elements of top ten right singular vectors for the total cytoplasmic DAPI intensity feature.

Figure 9:

Scatter plot (after removing ELN and NID from analysis) of elements of top four right singular vectors against each other for the cell compactness feature. Shape and color indicate ECOMP of the spot corresponding to the elements of the singular vector. The clusters seen in the NT panels are from missing spots on an outlier plate (see Suppl. Figure 31)

Figure 10:

(A) Scatter plot of elements of top four right ASVs calculated over 18 features measured on all MEMAs. Shape and color indicate ECMP of the spot corresponding to the elements of the singular vector. (B) Heat-map of top ten right ASVs calculated over 18 features measured on all MEMAs.