Contributions to Statistical Image Analysis for High Content Screening

by

Fangyi Liu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2012

Doctoral Committee:

Professor Kerby Shedden, Chair
Professor George Michailidis
Associate Professor Liza Levina
Associate Professor Gus Rosania

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Contributions to Statistical Image Analysis for High Content Screening

by

Fangyi Liu

Chair: Kerby Shedden

Images of cells incubated with fluorescent small molecule probes can be used to infer where the compounds distribute within cells. Identifying the spatial pattern of compound localization within each cell is very important problem for which adequate statistical methods do not yet exist.

First, we asked whether a classifier for subcellular localization categories can be developed based on a training set of manually classified cells. Due to challenges of the images such as uneven field illumination, low resolution, high noise, variation in intensity and contrast, and cell to cell variability in probe distributions, we constructed texture features for contrast quantiles conditioning on intensities, and classifying on artificial cells with same marginal distribution but different conditional distribution supported that this conditioning approach is beneficial to distinguish different localization distributions. Using these conditional features, we obtained satisfactory performance in image classification, and performed to dimension reduction and data visualization.

As high content images are subject to several major forms of artifacts, we are interested in the implications of measurement errors and artifacts on our ability to

draw scientifically meaningful conclusions from high content images. Specifically, we considered three forms of artifacts: saturation, blurring and additive noise. For each type of artifacts, we artificially introduced larger amount, and aimed to understand the bias by 'Simulation Extrapolation' (SIMEX) method, applied to the measurement errors for pairwise centroid distances, the degree of eccentricity in the class-specific distributions, and the angles between the dominant axes of variability for different categories.

Finally, we briefly considered the analysis of time-point images. Small molecule studies will be more focused. Specifically, we consider the evolving patterns of subcellular staining from the moment that a compound is introduced into the cell culture medium, to the point that steady state distribution is reached. We construct the degree to which the subcellular staining pattern is concentrated in or near the nucleus as the features of timecourse data set, and aim to determine whether different compounds accumulate in different regions at different times, as characterized in terms of their position in the cell relative to the nucleus.

CHAPTER I

Introduction

1.1 High content screening

High Content Screening (HCS) is a high-throughput experimental technique that is used to study living cells. In HCS experiments, an image of each experimental well is collected and stored for subsequent analysis. In contrast, traditional high-throughput screening collects only a small number of scalar summaries for each experimental well. HCS is now widely used for a variety of applications in biological research, such as implementation of systematic genome-wide functional screens [10].

In a typical HCS experiment, live cells are maintained in the wells of a 96 well or 384 well micro-titre plate. Cells in different wells may be experimental replicates, or may differ according to one or more factors of interest. For example, HCS is commonly used to study the effects of small molecule probes (or compounds) on cells. In such studies, a robot pipettes different compounds, or different concentrations of a compound into the different wells on a plate. After an incubation period, a digital image of each well is obtained using a magnifying objective. Moderate-sized experiments can generate hundreds of thousands of images in days. Therefore, automated image processing and analysis are crucial for deriving meaningful scientific conclusions from these studies.

In this thesis, we develop novel statistical approaches to handle data analysis

questions that arise when working with HCS data. A large community of researchers are developing image analysis methods for biological images, including images of live cells, using techniques from signal and image processing, mathematics, and computer science. We focus here on several problems that have received minimal attention to date, and that are particularly statistical in terms of their statement, and in terms of our approach.

In Chapter 2, we consider a particular construction of image features that is motivated by considering the relationship between contrast and intensity in HCS images. We show that these features are informative for an image classification task, and can also be used to reduce the dimension of the images for visualization. In chapter 3, we consider measurement error in HCS images, focusing on how the presence of measurement error biases downstream results. We identify situations in which this bias can, and cannot be effectively reduced, and we develop bias correction procedures that are effective in relevant settings. Chapter 4 considers the problem of analyzing timecourses of HCS images. Here the goal is to identify particular attributes of the images that vary across time in an interpretable way. We develop an approach for analyzing co-labeled image timecourses that can extract information about changing patterns of cell-associated, signal without requiring that the cells be segmented or tracked over time. Chapter 5 proposes some directions for future statistical research in this area.

1.2 Studies of subcellular localization

When a live cell is incubated with a small molecule probe (or compound), the compound may or may not enter the cell, and if it does enter the cell, it may localize to specific compartments within the cell, or it may localize diffusely within the cell. The behavior of a compound in a cell is determined by the chemical properties of the compound and the biological properties of the cell, and is influenced by other

factors such as the environmental conditions in which the cells are living, and the dose and incubation time of the compound. The nature of these relationships appears to be complex, as compounds with similar chemical structures can have very different localization behaviors. We are particularly interested in the following questions:

- Dose a specific chemical structure have a characteristic distribution pattern in most or even all cells?
- Do chemical substructures exist that confer localization tendencies to a broad class of compounds containing the substructure?
- Do substantial trends exist between physical and chemical properties of the compounds, and their localization behavior in cells?
- How much of the heterogeneity in the responses of the cells can be attributed to identifiable baseline properties of the cells?

In attempting to understand how small molecules behave in cells, a number of difficult questions arise. The images contain widely varying numbers of cells, the cells are irregularly distributed in the images, and the cells are often stressed by the conditions of the experiment. The compounds have widely varying optical properties and solubilities. For image analysis, we must deal with challenges such as uneven field illumination, high noise, low contrast, poor focus, limited dynamic range and numerous sources of artifacts including debris and precipitation resulting from poor solubility of the compounds.

High content screening is well-suited for studying phenomena relating to the patterns and dynamics of how small and large molecules distribute within cells. Pioneering work on localization patterns of proteins was begun in the late 1990's by Dr. Robert Murphy at Carnegie-Mellon University [5] [37]. Using established genetic transfection methods, Murphy's group labeled specific cellular proteins with green-fluorescent protein (GFP). Using large-scale imaging techniques, they then collected

thousands of images on dozens of proteins, providing detailed information about the variation of protein distribution patterns within and between protein classes. The use of GFP, or highly-specific optical probes, allowed Murphy's group to obtain very high quality images, with high contrast, high resolution, and low noise. Although the data were of very high quality, it nevertheless was necessary for this group to develop or adapt a number of image analysis techniques to allow meaningful conclusions to be drawn from such large collections of digital images. For example, Murphy's group has pioneered the use of machine learning techniques including support vector machines (SVM) and boosting for image classification [28], has extended these methods to accommodate overlapping localization patterns, and has developed original texture synthesis approaches for describing complex patterns of subcellular localization.

The number of proteins and the diversity of their localization patterns is relatively small – there are at most tens of thousands of protein coding genes in humans, and a much lower diversity of distinct subcellular distribution patterns. In contrast, the number and diversity of small molecules is essentially infinite, particularly if we consider endogenous small molecules, as well as exogenous molecules that may be introduced as drugs, or encountered in the environment. Moreover, there is no experimental technique such as GFP labeling that can be used to systematically label an arbitrary collection of small molecules for imaging studies. Most imaging studies focusing on small molecule rely on intrinsic fluorescence resulting from the chemical's structure, which is often weak. Therefore HCS imaging studies involving small molecules tend to have lower resolution, lower contrast, and greater noise compared to imaging studies involving fluorescent-labeled proteins.

In spite of the technical challenges, major scientific value would result from an improved understanding of small molecule distribution patterns in cells. For the development of therapeutic drugs, it would be valuable to understand where drugs localize within cells, since localization in sites remote from the intended target can

lead to side effects, and poor localization at the site of the target can lead to poor efficacy. Ultimately, it would be desirable to design drugs that not only have high affinity for their target, but also accumulate specifically in the regions of the cell where the target is located.

Outside the context of drug development, there are numerous other ways in which understanding subcellular localization patterns of small molecules can be scientifically valuable. Environmental chemicals such as pesticides are known to accumulate in certain organs and tissues, but little is known about their subcellular distribution properties. Understanding where these compounds accumulate within cells may provide important information about which environmental compounds are, and are not, health risks at low concentrations. Understanding the localization properties of metabolites and other endogenous small molecules may lead to a better understanding of cellular metabolism, with implications for treating health conditions such as obesity, diabetes, and cancer.

1.3 Image collection and analysis

Images are two-dimensional or three-dimensional pictures, which are obtained by using optical instrument to reflect certain appearance of some objects. Based on different purpose of image analysis or different sources of images, two-dimensional and three-dimensional images are used in different types of areas. For example, in the field of character recognition [7] or document image analysis [20], two-dimensional images are widely used, and in public health research, such as the analysis of brain images [50], three-dimensional images represent a large proportion in all images.

The technique of optical instruments can largely determine the quality of images. In modern sciences, obtaining high-quality, high-dimension (3D) live images with high-speed are important to answer problems in many areas. The better the image, the more information from the images can be analyzed and used [8].

Image analysis is the procedure of exacting valuable information from images, it has wide applications ranging from astronomy, geography, photography, public health research [16] to psychophysics. Historically, in the field of interpretation of fluorescence microscope images, the localization analysis was carried on manually [45], which was not only very cumbrous and time consuming but also had tight limitation, such as that it made the analysis of large amount of data set impossible. In the recent years, automated analysis developed a lot. People have investigated many approaches to interpret microscope images automatically, of which one elementary and critical idea is feature extraction, such as morphological features which are object-based [39], location features [4] and statistical features [45, 18]. Different models for image analysis have developed a lot as well, such as active basis model [] wavelet sparse coding [35] [40] [9] and Markov random fields [3] [21] [22] for image modeling and representation [47].

There are many other successful applications of statistical image analysis in science and engineering.

- Terrain recognition

  Terrain recognition is usually to study geomorphologic schemes, which can exist ranging from simple pathfinding to advanced area decomposition, from obstacle detection to route selection, and geographic categories matching as well [1] [32] [30]. It involves the symbolization and representation of a landscape, which may be 'mountain', 'plain', 'valley' and etc [32]. The features used in terrain recognition are usually the geographical features, which include contour lines, spot heights, depth values, skeletal lines and etc. [44]

- Face recognition

  Face recognition is very popular and attractive in recent years due to its wide applications in security, law enforcement and human-computer interactions [34]. A general process of face recognition could be stated as to automatically identify

or verify one or more persons by recognition of the faces, after still or video images of a scene are given [19]. One important attribute is that the recognition based on both still frames and videos. Compared with still frames, video-based face recognition is a relative new field on account of the development of novel technology. Evidence accumulation over multiple frames can provide better face recognition performance, but the low resolution is its main concern [41].

- Medical imaging

  Medical imaging includes the techniques used to create images of the human body (or part of) for clinical and diagnostic purposes and the analysis followed up. Because of the applications for health care, the images produced by the specific techniques such as X-rays and ultrasound have very high resolution, which means that the demands for image processing, compression, storage and retrieval are higher than in almost any other application domains [2]. Therefore, the main focus of medical imaging is more on application of high-tech equipments rather than constructing statistical models.

1.4 Image data sets

The raw data set, which consists of 20,736 images of small molecules, comes from Gus Rosania's lab, in the School of Pharmacy, University of Michigan. The images were acquired using the Cellomics KineticScan HCS instrument located at the Whitehead Institute.

The chemical probe library analyzed in this data set consists of 1344 fluorescent styryl small molecules. Each of these compounds was synthesized by conjugating one of 8 pyridinium groups (A-H) to one of 168 aldehyde groups (1-168) ($8 \times 168 = 1344$). Compounds were incubated with cells and were imaged on 96 well plates. The 96 wells on each plate were organized with the 8 rows corresponding pyridiniums A through

H, and columns 2 through 11 of a plate corresponding 10 of the 168 aldehydes. Column 1 and column 12 contained commercial (non-styryl) control probes localizing to mitochondria. In all, compounds were imaged on 18 plates.

The images are acquired in the presence of a quenching compound in the cell growth medium, to minimize fluorescence signal arising from probe that has not entered the cells. HeLa cells were used as the biological target of the probes. A probe called Hoechst dye, which stains DNA and is highly selective for the cell nuclei and which usually fluoresces in a separate channel from the styryls, was added to each plate to allow easier identification of the cell nuclei. A marker for the cytoplasmic region or whole cell was not used, so the boundaries of the cell must be inferred (when possible), from the pattern of styryl fluorescence. Each plate was imaged under influx conditions (with styryl probe in the medium) and efflux conditions (after washing and replacing with fresh medium). The instrument acquired images simultaneously in four spectral channels, Hoechst($\sim 461nm$), FITC($\sim 520nm$), TRITC($\sim 570nm$) and Cy5($\sim 670nm$).

For each plate, six images were obtained under both influx and efflux conditions (12 images in all per plate). These six images are: Hoechst channel at 1s exposure; TRITC, FITC, and Cy5 channels at 1s exposure; and TRITC and FITC channels at 200ms exposure. Each one of the digital images are 512×512 pixels in size, with 12 bit intensity values ranging from 0 to 4095. The fluorescence intensity is approximately linear with the concentration of probe, however, saturation can occur, yielding intensities of 4095.

As for reference, a publicly available collection of images was analyzed as well [26]. This data set consists a set of 862 fluorescence microscope images of HeLa cells[45, 6, 12]. Different protein labels were used in the reference data set, and based on the known localization of the labels, ten distribution patterns were defined as: actin filaments, endoplasmic reticulum, endosomes, Golgi marked with giantin, Golgi marked

with GPP130, lysosomes, microtubules, mitochondria, nucleolus, and nucleus.

CHAPTER II

Conditional Contrast Patterns as Features for Image Analysis of Cells

In this chapter, we develop a set of easily-computed image features based on the intensity contrast stratified over the intensity magnitude in spatial neighborhoods of various sizes, to allow the relevant parts of the intensity range to be adaptively identified in each image. These image features are used to accurately predicted expert-assigned classifications of staining patterns, as part of analyzing fluorescent molecules with diverse localization patterns. The necessity of conditioning is demonstrated in this chapter as well, in a way of exploring artificial cells with different degree of spatial dependence and several real data set as for reference. In the final part of this chapter, we perform image classification analysis of weakly fluorescent molecules with diverse localization patterns, and its regularization and assessment of sensitivity to feature parameters as application of the conditional contrast model. By comparing performances based on different spatial scales, the results indicate that at least two distinct spatial scales must be modeled to achieve good classification performance in both the small molecule data set and our reference data set from a location proteomics experiment using green fluorescent protein (GFP) to label cellular substructures.

2.1 Feature construction

2.1.1 Texture-based and object-based features

Feature extraction is an important step in image analysis, it reduces dimensions of data by transforming the input data into a set of numerical features. If the features are selected well, we could expect that they are representative of the meaningful information contained by the raw data. Texture-based and object-based features are two fundamental elements used in numeric interpretation of images.

It is difficult to explicitly define what is 'texture' in a mathematical way, and there is no generally agreed upon definition. However, people have some common notions that the texture of an image refers to the spatial distribution of gray tones, the structural organization of surfaces and relationship among their neighborhoods [24].

Texture-based features are fundamental characteristics of images, statistical texture-based features usually are a common set of general statistical parameters usually include contrast, uniformity, and homogeneity [24]; model-based features could be generated by Markov random field(MRF) models [13] [48]; wavelet-based features could be yielded by Gabor filters or ICA filters [46].

Unlike texture-based analysis, most object-based analysis relies on the assumption of within-region localization-independence of the features, in other words, the regions should be homogeneous with respect to some characteristics such as color and texture [27].

Due to the factors such as uneven field illumination, low resolution and noise, variation in intensity and contrast, and cell to cell variability in probe distributions, we hypothesized that a texture-based analysis would be competitive with a more traditional object-based analysis in our specific application.

Another important issue arises in account of the variation: the scaling. Intensity scaling and size scaling are two major aspects people are concerned with in current re-

search topics. Images with similar patterns but different brightness may be considered as different categories of images in other discipline.

However, when cells were incubated with chemical probes, the intensity would not reflect to localization distribution very much, therefore, intensity scaling is not our primary concern. We are not interested with the size scaling, since some specific probes may have the tendency to enter larger organelle such as mitochondria.

2.1.2 Gabor filters

A standard approach to texture analysis is to convolve the image with a filter in two dimensions, then extract numerical summaries using the pixel intensities in the resulting filtered images. The filters we used are transformations of Gabor filters [38], which have a variety application in image analysis [33] [29] [36].

The general version of two-dimensional Gabor filter function is defined as

$$G(x, y; \lambda, \theta, \phi, \sigma, \gamma) = \exp(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}) \cos(2\pi \frac{x'}{\lambda} + \phi),$$

where $x' = x\cos(\theta) + y\sin(\theta)$ and $y' = -x\sin(\theta) + y\cos(\theta)$. In this equation, λ represents the wavelength of the cosine factor, θ is the angular parameter that determines the rotation of the normal to the parallel stripes of a Gabor function, ϕ is the phase offset, σ is the scale of the Gaussian envelope and γ is the spatial aspect ratio, which controls the ellipticity and orientation of the filter peaks.

We used the standardized version of Gabor filter functions $F = a + bG$, where a and b are scalars so that

$$\int\int F(x, y; \cdot)\, dx\, dy = 0 \text{ and } \int\int |F(x, y; \cdot)|\, dx\, dy = 1.$$

By choosing different values of parameters, different shapes and spatial scales are emphasized. We pre-specified filter parameter values without optimizing the

performance for classification or visualization. We chose $\phi = 0$ and $\lambda = m = 2\sigma$ so that the shape of F has two local extreme values within the bounds of the filter's footprint, one is local maximum and the other is local minimum, and there is an anti-symmetry in which rotation by 90 degrees following by multiplication with -1 returns the original filter. Based on prior knowledge that the average size of an individual cell is about 80 pixel diameter, and the cells have diverse shapes, we pre-specified baseline parameter values as follows: γ in (1, 5, 10) and m in (5, 10, 20). In addition, considering the features need to be invariant to rotations since there is no regular pose for the cells, we pooled the convolved value together after filtered by four versions, for each filter function with the same other parameter values (shown in Figure 2.1): using θ in $(0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4})$.

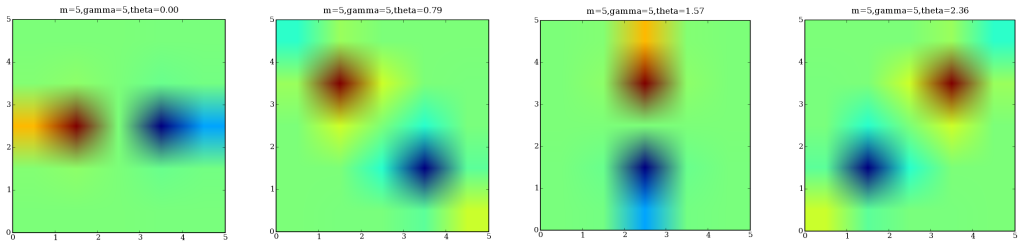


Figure 2.1: Four rotations of Gabor filters.

Figure 2.2 gives three examples, with one circular filter function, one vertically elongated filter function and one horizontally elongated. Given an authentic example of single cell shown in the left part of Figure 2.3, the right part of Figure 2.3 shows the result of how the circular filter (the left one in Figure 2.2) works.

2.1.3 Features based on contrast stratified over intensity

As the properties of Gabor functions defined in Subsection 2.1.2, the filter F is L1 normalized and has mean 0 so it represents contrast values, and $|F|$ represents the corresponding intensity values, as shown in Figure 2.4.

Therefore, if we use IM to present an image and F denote a filter, then $|IM * F|$ gives the contrast convolved image, and $IM * |F|$ gives the intensity convolved

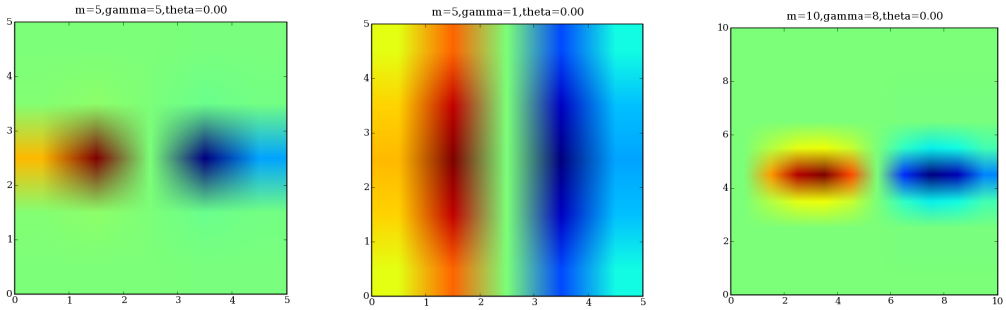


Figure 2.2: Examples of Gabor filters, from left to right are circular, vertically elongated and horizontally elongated filters.

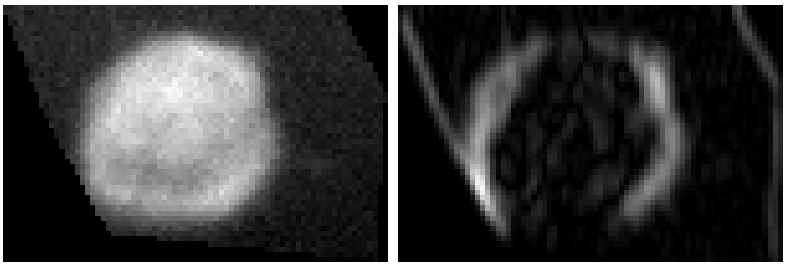


Figure 2.3: From left to right: an authentic example of round cell; the corresponding filtered version after filtering by the left filter in Figure 2.2.

image, where $*$ represents convolution in two-dimensions. The pixels of $|IM * F|$ and $IM * |F|$ are paired based on their positions in the image. This procedure could be demonstrated as follows.

Given a round cell as shown in the left part of Figure 2.5, the middle picture of Figure 2.5 shows the 'contrast""($|IM * F|$) convolution after filtering by the left filter in Figure 2.4. Through the comparison of these two images in Figure 2.5, we could see that within the contrast version, the boundary of the cell has been exaggerated because the levels of pixels inside and outside the cell have obvious difference, while the interior of the cell has been eliminated a lot due to the fact that pixel values in it are similar. The right part of Figure 2.5 gives the 'intensity' ($IM * |F|$) after filtering by the right part of Figure 2.4, which gives a smoother version of the original image.

Considering images of both intensity and contrast of the same cell contain certain information of the localization pattern, however, different patterns may give the same

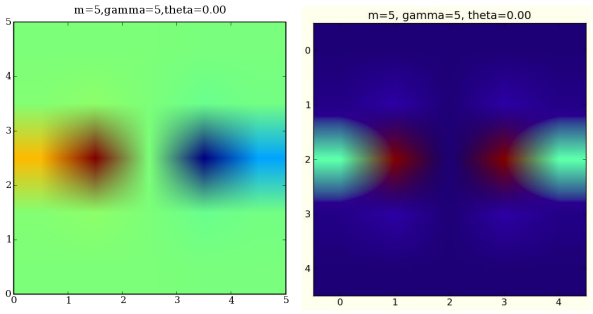


Figure 2.4: From left to right: one circular of Gabor filter (F); the corresponding absolute filter function ($|F|$).

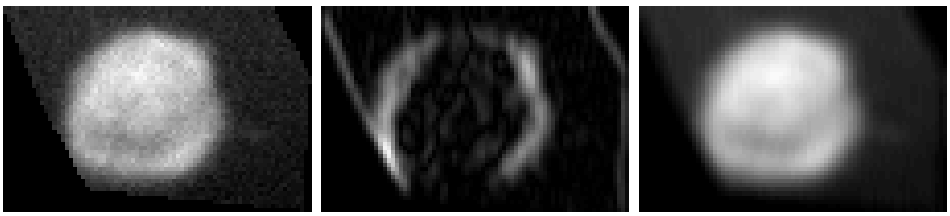


Figure 2.5: From left to right: the intensity image of left part of Figure 2.3; filter function of two positive parts.

marginal distribution of intensity or contrast, and many contrast patterns may be more or less informative based on the local brightness where the pattern occurs, that's the reason why we want to study the conditional relationship between intensity and contrast, and we want to condition on 'intensity' specifically. This will be explored more in Section 2.2.

As the construction of initial features, for a given rectangular box (refer Section 2.3) of a single cell, we applied the linear stretch($\frac{IM-\min(IM)}{\max(IM)-\min(IM)}$) on the image IM first, and then produced the 'intensity' version named as I and the 'contrast' version named as C accordingly. We paired the (C, I) versions by pixel location, and pooled over the four values of θ, which is described in the selection of parameter values in Subsection 2.1.2. Thereafter, we stratified the (C, I) pixel pairs into three parts based on the value of the intensity magnitude, where the cut points are the marginal median and 90th percentile of I. Within each of these three pieces, we calculated four quartiles of C (the 25th, 50th, 75th, and 90th percentiles) as image

features. This gives us 3(strata)*4(quantiles)=12 image features per filter/per image, and these features are constructed separately for each of 9 filters ((3 values for γ) * (3 values for λ) = 9 filters). Therefore we have 9*12 = 108 numerical features for each image. In addition, since the small molecules were co-localized with Hoechst dye (refer Section 1.4), this can be done both for the whole cell image, and for the nucleus (based on thresholding of the Hoechst channel to identify the nuclear regions, which is demonstrated in Section 2.3), thus, a total set of 216 features have been obtained ultimately per individual cell image for the small molecule data set. The protein images were not co-localized with Hoechst dye, so we used only the 108 whole-cell features when analyzing the protein data set.

Figure 2.6 shows two examples of the conditional features pattern of the cell image, the left one comes from the round cells of the small molecule data set, and the right one comes from the autofluorescent category of the small molecule data set. For the conditional quartile features (refer Figure 2.6), the vertical partitioning is invariant with intensity scaling since the relative positions of the quartiles is unvarying, but the horizonal numerical values are depend on the intensity, the higher the intensity is, the larger the contrast quartiles are. The comparison of these two plots shows diversity of feature patterns among different categories.

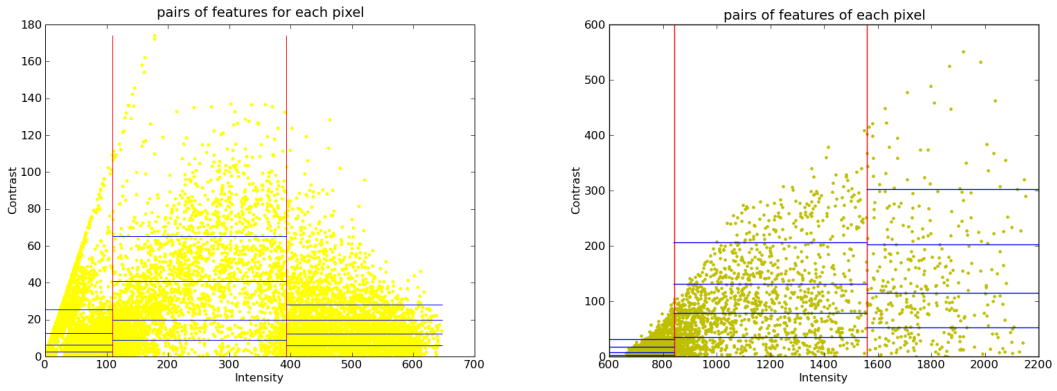


Figure 2.6: Two conditional feature patterns.

2.2 Statistical support of conditional features

2.2.1 Artificial cells

As demonstrated in Subsection 2.1.3, we constructed image features by selecting the conditional model, i.e., features are contrast quantiles conditioning on intensity magnitude. In order to figure out whether the conditional method is necessary to obtain enough information from the images and how, compared with the marginal distribution of either contrast or intensity, we then again extracted the same four quantiles (the 25th, 50th, 75th, 90th percentiles) but only from the marginal contrast and intensity distribution respectively, and at this time, we only considered extracting features from the whole-cell region, thus for either way, we have 4(features)*9(filters) = 36 features per image. In addition, considering the difference of dimension between conditional model (108) and marginal model (36) is noticeable, we then simply combined the two marginal feature set, thus the new feature set has the dimension as 72.

We first applied the procedure to artificial cells. To simplify the question, we only considered the two way classification problem, which means that the cells generated only come from two different distributions. And we performed classification analysis by using the leave-one-out LDA. Considering that to construct cells similar as in real life is way too complicated and time-consuming, and various factors of real cells such as shape, rotation or size would not influence much on the cellular localization distribution by using texture-based features, therefore, for each category of artificial cells, we chose the same cell size and sample size. In addition, we constructed a circular pattern as nuclear region, the combination of this circle and an outer ring as whole-cell region, and we created the inner cell (nuclear) region and outer cell region (the outer ring) with approximately same area, as a simplified situation. Let $P_{i,j}^{k}$ denote the pixel value at position $(i,\ j)$ for class k, then $P_{i,j}^{k}$ comes from a normal

Table 2.1: Classification rates with standard deviation for artificial cells.

Sample size	Feature set			
	conditional	intensity	contrast	combination
Large	0.86 (0.023)	0.70 (0.028)	0.72 (0.031)	0.77 (0.030)
Moderate	0.80 (0.043)	0.67 (0.052)	0.70 (0.043)	0.72 (0.049)
Small	0.009 (0.012)	0.63 (0.072)	0.64 (0.078)	0.49 (0.094)

distribution, and in our design, we have

$$E(P^1_{i,j}) = E(P^2_{i,j}) \text{ and } Var(P^1_{i,j}) = Var(P^2_{i,j}).$$

One example of artificial cell is shown in the left side of Figure 2.7.

Thereafter, we mainly focused on the factor which may have essential influence on localization distribution in our hypothesis, which is the spatial dependence of pixels. Moreover, in order to confirm that the classification results were not due to randomness and was reproducible, we repeated the classification procedure with same parameters for a hundred times, and the results were obtained based on the mean classification rate. By using different parameter (mean pixel difference between inner and outer cellular region and spatial dependence degree) values, we figured out that the degree of spatial dependance is the major factor to differentiate two different distributions when we have moderate sample size. However, when the sample size decreased sharply, due to the fact that conditional set has 108 features while marginal set has only 36 features, the conditional model would not perform as good as marginal feature set. Such results are shown in Table 2.1, the conditioning model did give noticeable better performance on classification problem only if the sample size is not extremely small, thus our further analysis was based on moderate sample size.

Considering that we stratified the features from conditioning model into three subregions: low intensity region, middle intensity region and high intensity region, we extracted features from these three subregions based on conditioning model, and calculated the z-score of the two categories for each feature. Z-scores from the two

marginal models were computed as well. The sorted absolute values of z-scores were plotted in Figure 2.7.

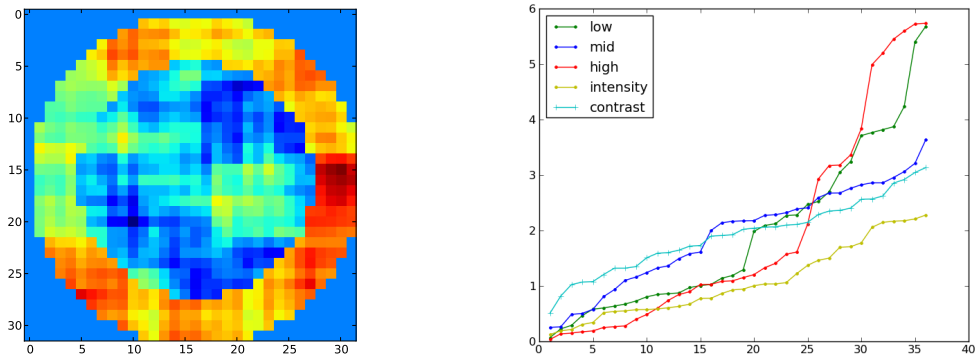


Figure 2.7: Left: one example of artificial cell. Right: sorted absolute z-scores for features from different models.

From the right plot of Figure 2.7, we would notice that the features of conditioning model from low and high intensity regions have highest (absolute) z-score values in general, and the overall z-scores from the intensity model generated the smallest (absolute) z-score. These results are consistent with both the classification rates, if we used features from low, middle, or high intensity region alone as classifiers, and the way we constructed the cells: due to that the areas of inner and outer cell region is almost the same, the inner and outer cell region located approximately the lower half and the upper half part of sorted intensity pixels, this explains why features from low or high intensity region would be more differentiated between two categories, compared with those features from mid intensity region. In addition, contrast filters gave approximately 0 values except for the boundaries, which indicates that the boundaries would contain some information, therefore, the classification based on features from marginal intensity model was not purely guessing, even though we don't know exactly what information could be obtained from the boundaries, but it indeed contains least information compared with other models.

To further understand the statistical reason behind, we intended to explore the

information contained at boundary. In the left plot of Figure 2.7, the boundary between inner cell region and outer cell region and the boundary between outer cell region and the background have different size, to further simplify the question, we designed an artificial cell pattern as shown in Figure 2.8, the circles in the left and right part represent the inner and outer cell region, and they have same area, therefore, the boundary length is the same as well. In the middle is the buffer area, so when the image is filtered, the two circle parts will not be influenced crossly. We still only consider the two-class situation. For class 1, one circle part has low mean intensity with low dependence, the other circle has high mean with high dependence. For class 2, one has low mean with high dependence, while the other has high mean with low dependence.

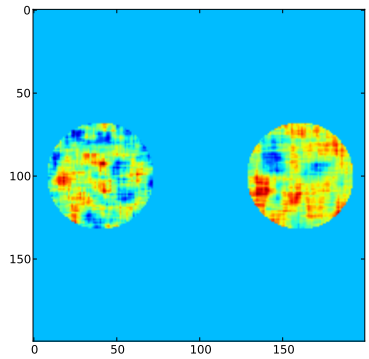


Figure 2.8: One example of the newly designed artificial cell pattern.

From the previous analysis, we would believe that independent features could contribute more when each feature contains same information as dependent ones. As a way to measure the degree of dependence, for each individual feature, we calculated the ratio of within-class variance to total variance, and sorted these features descendingly by the ratio, which is shown at the left part of Figure 2.9. We could learn that the overall trend is that the stratified (conditional) features are less depend with each other, relatively speaking, while the features come from marginal intensity field are highly related with each other, more related than features from marginal contrast

model.

Subsequently, we are going to look for how much unique information each feature contains, thus we generated a very large sample to estimated the population mean and variance, considered the non-cross-validated linear discriminant analysis of subset of features, by starting classifying with one feature which has the highest ratio, then add one more feature with next highest ratio into the subset of features each step, the results are shown on the right part of Figure 2.9. Learned from the plot, we could know that each time you added one more feature (which means more information), you would get better classification performance, but the speed of increment become slower when enough information were obtained, even you added more, the information was still fairly redundant, that is why the curve becomes more flattened when the number of features increases.

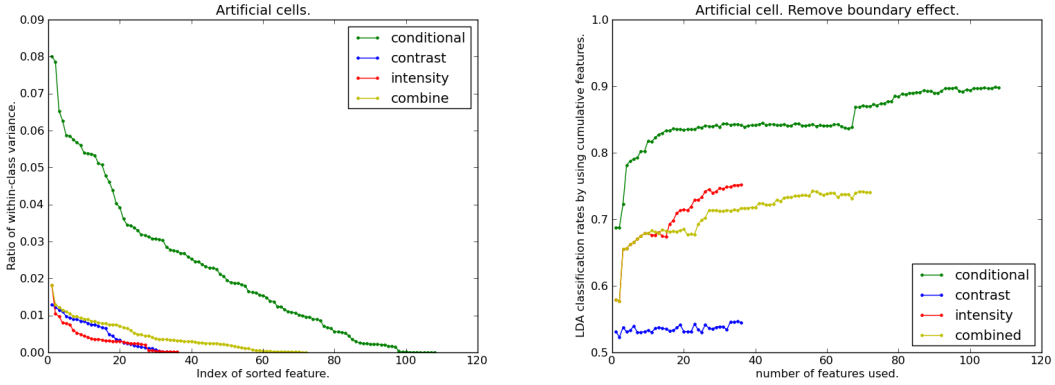


Figure 2.9: Left: ratio of within-class variance for each individual feature (sorted). Right: Estimated classification rates by using subset of features cumulatively.

2.2.2 Real cells

In addition, we applied the same procedure to real cells. For the protein data set, the conditional model gave a 89.68% (83.58%, 95.78%) overall classification rate, which was apparently a better performance with regard to classification problem, by

comparing with a 75.52% (72.48%, 78.56%) classification rate given by marginal intensity quantiles and a 82.60% (79.94%, 85.26%) classification rate given by marginal contrast features, also the simple combination of two marginal feature sets performed a 85.73% (83.25%, 88.21%) classification rate.

Another data set was obtained from Murphy's lab at [15], has a total of 327 fluorescence microscope images of Chinese hamster ovary (CHO) cells, and those were labeled with 5 different categories, the Golgi protein giantin, the lysosomal protein LAMP2, the yeast nucleolar protein NOP4, and tubulin (Sigma). By applying leave-one-out LDA to this data set, the classification rates with 95% confidence interval were 98.17% (88.29%, 100%) by using conditional feature set, 93.88% (91.54%, 96.22%) and 92.97% (89.21%,96.73%) by using marginal intensity and contrast feature set correspondingly, and 97.55% (83.91%,100%) by using the simply combination of the two marginal feature sets. Thus, by using conditional features was several percentiles better than using marginal features.

A third data set is the multi-cell yeast images which is now hosted by SGD and was originally designed and built by the laboratories of Erin O'Shea and Jonathan Weissman at the University of California, San Francisco (UCSF). This is used to study protein localization in the budding yeast [49]. In the UCSF yeast data set, 6029 open reading frames (ORFs) were GFP tagged, and each protein were assigned to one or mixtures of 22 unique location categories (with one as 'ambiguous'). For each set of images, there are three channels of the same field of yeast cells: a DAPI image shows the DNA distribution, the DIC image highlights cell boundaries, and a GFP image shows the location pattern of the tagged protein. In previous protein localization study, people used 2713 image set by removing those were signed to 'ambiguous' category and had multiple labels, without cell-level segmentation, the overall classification rates was 76.2% applied to the full 21 classes, which are cytoplasm, nucleus, mitochondrion, ER, vacuole, punctate_composite, nucleolus, cell_periphery, vacuo-

lar_membrane, nuclear_periphery, spindle_pole, endosome, late_Golgi, actin, peroxisome, lipid_particle, Golgi, bud_neck, early_Golgi, microtubule, ER_to_Golgi. [14].

In our project, to simplify the question, we only considered GFP images for each yeast cell, and we did not apply image segmentation as a consideration of efficacy because yeast cells are often clustered together. To simplify the question, we only considered the images with one unique label outside of 'ambiguous' category. Therefore, we have 2728 images for 21 categories, and among them, about 76% of the images (2071) belongs to four major categories, which are 'cytoplasm', 'nucleus', 'mitochondrion' and 'ER'.

At the very beginning, we only selected a subset of images. We chose a list of 150 images for each of ER, nucleus, mitochondrion and cytoplasm by arbitrary, and chose a full class of actin (28), microtubule (12), Golgi (42), nucleolus (64) and endosome (49), therefore, we had a subset of yeast data with 796 GFP images from 9 categories. Considering that the noise have a very large proportion of pixels, we threshed each image by using pixel value of 0.25, which is the 90th quantile on average. Therefore, the conditioning model had a 73.24% accuracy, while marginal intensity and contrast model gave 49.75% and 61.68% overall classification rates, respectively, and the combination of two marginal features performed 68.22% accurately. We then applied the same procedure to the whole data set with full 21 categories. It turned out that the conditional model classified 65.91% accurately, the marginal intensity and contrast model resulted 33.43% and 50.92% rates respectively, and the combined marginal feature set was with 60.67% accuracy. Considering that the largest 4 categories, which are cytoplasm, nucleus, mitochondrion and ER, contributed over three quarters of the whole image set, the classification rates based on conditional, intensity, contrast and combined marginal feature set were 87.01%, 70.93%, 81.07% and 86.09%, correspondingly. Based on these analysis, it was noticeably that the marginal features were not as persuasive as the conditional features. Therefore, to model the quantiles

conditionally was non-negligible in our feature construction method.

Considering the features of conditional model come from three different strata, region with low intensity, region with middle intensity, and region with high intensity, to explore the quantiles from these regions are redundant or not, we calculated the (absolute) z-scores for each pair of classes in yeast data, and plot the average z-scores increasingly. From left plot of Figure 2.10, we could see that differences exist between features under different models. While we plotted estimated classification rates by adding one feature each step as in Subsubsection 2.2.1, we could know that we still have space to add more informative features because the increment curve does not achieve its asymptotic results yet.

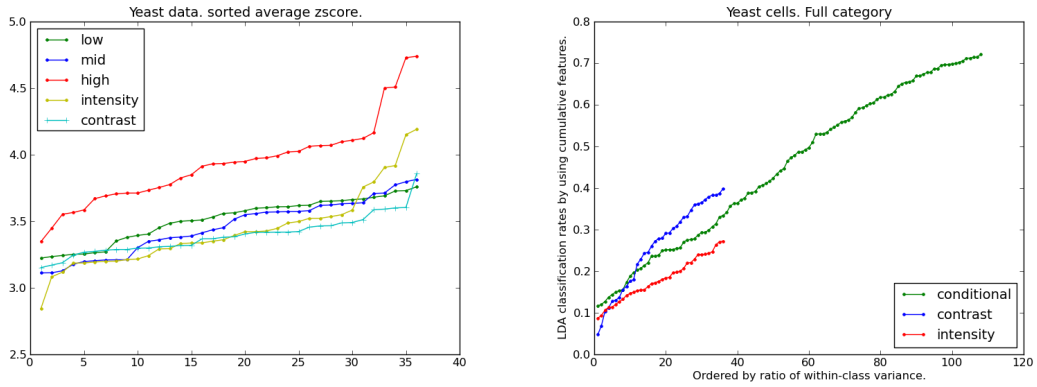


Figure 2.10: Left: The average of pairwise z-scores for yeast data (sorted). Right: Estimated classification rates by using subset of features cumulatively.

As a study of dependence between pixels in real data set, we chose the protein data set, and measured the expected absolute difference of pixels in a neighborhood when those pixels all came from the same intensity subregion. From Figure 2.11, we could see that different categories have different relationship between pixel mean and pixel covariance. This relationship comes from the raw image, which provides the evidence that the stratified feature model could capture such information in a certain range and get better performance with regard to classifying localization intensity distribution.

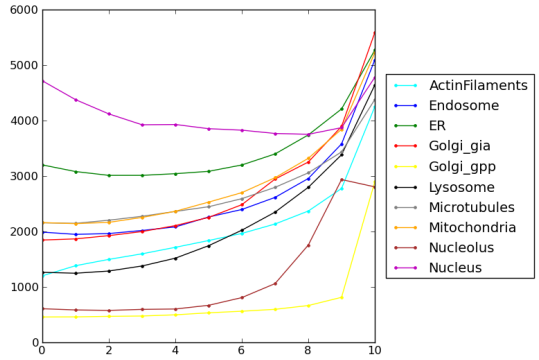


Figure 2.11: Dependence measurement for the protein data set.

2.3 Image pre-processing

In the raw data set, since Hoechst dye was added to each plate, it is possible for us to identify the cell nuclei regions by thresholding the Hoechst channel to obtain a nuclear mask. Figure 2.12 shows six representative examples of raw Hoechst channel images. It is apparent that there is substantial variation in the number, size, and positioning of cells in the images, and in the intensity and contrast of the images. Due to the variation of brightness of images and the contrast between the interior and exterior of the cells, and the fact that threshold values must be dependent on the brightness, we need to identify the proper threshold values for each image.

The idea we followed was to pick a thresholding value to make the masked image have the greatest number of distinct objects with proper object size, which we set the range from 100 to 800 pixels. For computational reasons, it is impractical to test all integers to figure out the ideal threshold, therefore, we started with an initial threshold value (800), then raised or lowered the threshold over a sequence of iterations to identify an approximately optimal threshold.

The position and angle of the light source influences the brightness of the cells and the background. Ideally, the light detected at a particular pixel is proportional to the amount of probe in the cell in a z-axis column above the pixel (ignoring focusing

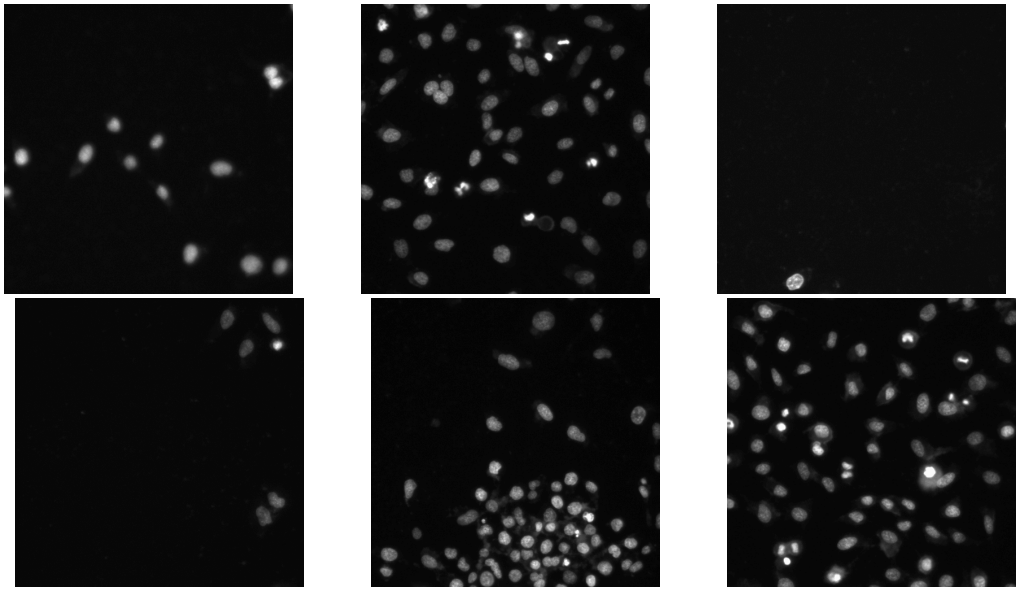


Figure 2.12: Several examples of nucleus Hoechst channel images.

effects). Therefore, uneven illumination in the field of view will prevent accurate quantification of probe abundances using images.

In order to correct the uneven illumination, and in view of large size ($512 \times 512 = 262144$) of the full image, we selected a random sample of the pixel values with size 5000, and due to the natural of lighting curve, we constructed a quadratic model with 6 degrees of freedom after log transformation in such a way: $\log(f) \sim x + y + x^2 + y^2 + xy$. Considering the existence of noise, which is treated as the outlier in the statistical view, we implemented robust regression by using the function lqs in R software. One example is given bellow to show the nice performance. The left image in Figure 2.13 gives a raw image with a bunch of cells, the middle one gives the image after thresholding by a certain value (the threshold value is not the same as the one to isolate cell nuclei as described in the beginning of Section 2.3), and the right image is the according pattern of fitted illumination function based on the quadratic model.

As a preparation of further analysis, we need to separate each individual cell, at first, we investigated one way to implement accurate cell segmentation. The concrete steps were: after nuclei region masked and background corrected, we attempted to

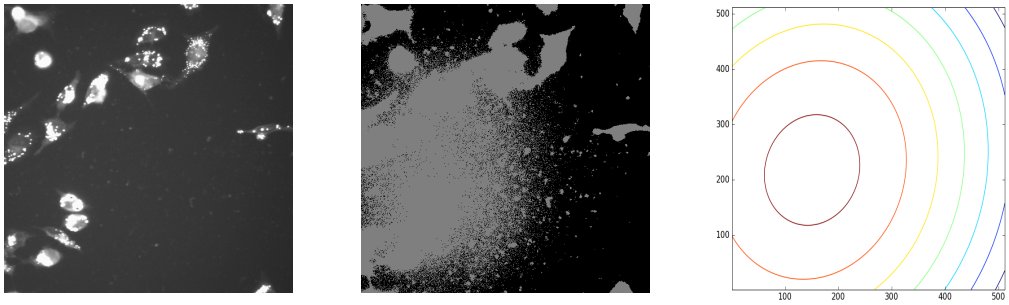


Figure 2.13: From left to right, the three images are a raw cell image, corresponding thresholded image and the fitted illumination function by the quadratic model.

extend the edge of cells one pixel by pixel, the criterion to judge whether a point at a specific position to extend or not is whether it is larger than the median intensity of the 1-dilated cytoring or not. After several iterations, we would obtain an approximate shape of each cell. The segmentation algorithm is too complex, tedious and time-consuming. In addition, through Figure 2.12 and Figure 2.13 we could see how variable in the number, size and shape the cells could be, thus the variation largely limited our effort even though we did a lot of work.

Therefore, for realistic purpose, we presented the following 'box' procedure:

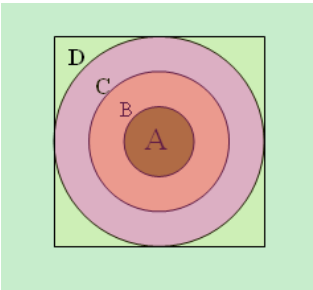
We masked the nuclei regions by thresholding Hoechst channel and did 20-pixel dilation, then we draw such a box around each dilated region that it is the smallest box which contains the dilated region.

This procedure did not give the accurate shape or segmentation of cells, however, it could save a large amount of time, and basically, the signal in one box comes from one single cell. Even if there are half-cells at the corner of a few boxes, it is worth-while applying this 'box' procedure rather than cell segmentation.

Many styryl compounds will be non-fluorescent, or will be impermeable to the cells. In these cases, no signal should be detectable from the probe, and all light defected in image comes from other sources such as background or cellular autoflu-

orescence. High background, low resolution, dense noise, and little signal all may make the image hard to analyze, those hardly analyzable images should be excluded in order not to influence the outcomes. We have the premise that the images with analyzable signal should have the property that, the intensity concentrates much more around the nuclei, rather than being distributed uniformly in the image. Therefore, our image selection was based on the basic statistics of the intensity, which are the ratio and difference of the intensity inside and outside the cells, and the proportion of saturated pixels (pixel with intensity 4095) in the cells. The procedure is:

By using the threshold values demonstrated in previous paragraphs, we masked the nuclei region with proper size at first, and then we did both 10-pixel and 20-pixel dilation around the nuclei region. Afterwards, the 'inside intensity' II was defined as the median intensity of the 10-pixel dilation region, and the 'outside intensity' OI was defined as the median intensity of the complement of the 20-pixel dilation region. Let $D = II - OI$ and $R = II/OI$, then D and R represent the intensity difference and intensity ratio respectively. We defined S as the proportion of pixels in the 20 pixel dilated regions that are at the peak level (4095), which is a measure of saturation.



As shown in the left picture, if A denotes the nuclear region and D is the box mentioned in previous paragraphs, then B (including A) shows the part after 10-pixel dilation and C (including B) is the region after 20-pixel dilation.

Via testing on training data (information about training data is on Subsection 2.3.1), we picked the cut points as: an analyzable image would be selected only if its $D > 300$, $R > 1.2$ and $S < 0.05$.

2.3.1 Training set

In the small molecule data set, 897 single cells have been classified into eight categories(autofluorescent, mitochondrial control, round, round mitotic, cytoplasmic membrane, mitolight, plasma membrane, and RNA) manually, and these 897 cells which are individual cells were cut out as polygonal sub-images from the full images. We are able to retrospect these cells into the original data set, in another word, we are able to associate each training set cell with a rectangular region in a specific raw image, which allow us to obtain raw data for all six channels for each training set cell, including Hoechst channel. As demonstrated in the Subsection 2.1.3, we constructed 216 features for each individual cell, which means we transformed the information from images into proper quantitative data.

For the protein data set, as described in Section 1.4, each image contains a single cell, so we do not need to apply the separation procedure. In addition, all the images were classified into ten categories, thus we used this whole data set as a reference training set.

Due to the fact that the imaging magnifications of the proteins and the small molecules are different, we manually measured the length of the major axis of the nuclei of 100 randomly selected cell images from each of the two image collections and calculated empirically, and it turned out that the proteins were imaged at higher magnification with a factor of 2.4.

This following analysis thereby was based on the conditional features of 897 small molecule images and 862 protein images.

2.4 Classification analysis

Image classification is the process to convert the spectral data automatically into several categories through image features, such as shape, brightness, orientation, cen-

ter of gravity of objects or some more complicated attributes. Two main modes of image classification are: supervised and unsupervised. In general, a supervised classification needs a prespecified fixed number of classes (groups) to assign each image into one of those classes, while unsupervised classification does not give labeled examples.

The image features were used to classify each image using linear discriminant analysis (LDA). Leave-one-out cross validation was used to unbiasedly estimate the classification rates, and the entire cross-validation procedure was bootstrapped 1000 times to provide standard errors for the estimated rates of correct classification.

By reason of the factor of imaging magnifications of the two data sets is between 2 and 3 (as in Subsection 2.3.1), and we are looking for the same resolution magnification for both data set, considering the computing efficacy as well, therefore, we down-sampled the protein data set by the factor of 2*2 and 3*3 accordingly and applied LDA. The classification rate (with 95% confidence interval) under baseline models by applying leave-one-out LDA was 73.3% (71.6%, 75.0%) for the full (8 class) small molecule data set, and the classification matrix is shown in Table 2.2. For a reduced small molecule data set based on four categories (cellular autofluoresence, mitotracker, plasma membrane, and RNA), the rate of correct classification was 86.5%(80.1%,92.9%). The classification rates for the full (10 class) original protein data set, down-sampled by 2*2 and by 3*3 were 89.68%(83.58%,95.78%), 88.52%(81.4%, 95.6%) and 88.4%(83.64%,93.16%), accordingly. Because these results are included in the confidence intervals pairwise, therefore, the further analysis based on the reference data set were talking for the down-sampled by 2*2 version of protein images, if not specified. Table 2.3 shows the classification matrix for the down-sampled by 2*2 protein data set.

		Predicted							
		AutoF	Mitotracker	Round	Mitosis	CMem	Mitolight	PMem	RNA
Actual	AutoF	**0.89**	0	0	0.03	0	0.08	0	0
	Mitotracker	0	**0.83**	0.14	0	0.06	0.01	0.08	0
	Round	0	0	**0.71**	0.13	0.03	0.02	0.02	0.10
	Mitosis	0.05	0	0.17	**0.59**	0	0.03	0.15	0.01
	CMem	0	0.06	0	0.02	**0.59**	0.11	0.18	0.05
	Mitolight	0.04	0.10	0.01	0.03	0.03	**0.76**	0.03	0
	PMem	0.01	0.07	0.03	0.05	0.10	0.06	**0.67**	0
	RNA	0	0	0.083	0.03	0.06	0.03	0.02	**0.78**

Table 2.2: Confusion Matrix by LDA(0.73) for the small molecule data set.

		Predicted									
		Actin	ER	Endosome	Golgi_gia	Golgi_gpp	Lysosome	Microtubules	Mitochondria	Nucleolus	Nucleus
Actual	Actin	**0.98**	0	0	0	0.01	0	0	0	0.01	0
	ER	0	**0.94**	0.01	0	0	0	0.02	0.01	0	0.01
	Endosome	0	0.01	**0.82**	0.02	0	0.11	0	0.03	0	0
	Golgi_gia	0	0	0	**0.80**	0.20	0	0	0	0	0
	Glogi_gpp	0	0	0.01	0.21	**0.76**	0	0	0	0.01	0
	Lysosome	0	0	0.07	0.04	0	**0.89**	0	0	0	0
	Microtubules	0	0.01	0.03	0	0.03	0	**0.89**	0.03	0	0
	Mitochondira	0	0.04	0.04	0	0.03	0.01	0	**0.88**	0	0
	Nucleolus	0	0	0	0.01	0.11	0	0	0	**0.88**	0
	Nucleus	0	0.01	0	0	0	0	0	0	0	**0.99**

Table 2.3: Confusion Matrix by LDA(0.89) for the protein data set down-sampled by 2*2.

2.4.1 Spatial scale factor

Our baseline feature set uses filters corresponding roughly to 6%, 13%, and 25% of the typical cell diameter (corresponding to m=5,10,20). The spatial scales of the objects of interest ranged substantially – for example, RNA, mitochondria, and other organelles have a much smaller spatial scale than the cell nucleus and plasma membrane. Thus we were interested in assessing whether image features with multiple spatial scales must be included in the feature set to be able to accurately classify the images. We repeated the classification analysis using one, two, and three different spatial scale values, optimizing the spatial scale values to maximize the classification rate. For the small molecule data set, dropping from three separate spatial scales in the baseline model (pre-specified without optimization) to two spatial scales (optimized to maximize performance) only resulted in a 0.5 percentage point drop in overall classification rate. However using only one spatial scale caused a 5.1 percentage point drop in classification rate compared to using two distinct spatial scales. Thus we concluded that the small molecule data can be adequately described using features at two spatial scales, but one spatial scale is insufficient. In the protein data set, we applied the analysis to the images with original imaging magnification (without any down-sampling), the drops in classification rates were 1.5 percentage points and 5.8 percentage points drop when comparing three scales to two scales, and when comparing two scales to one scale, respectively. Thus the necessity to model features at two spatial scales was a common attribute of these two very different image collections.

Figure 2.14 shows the classification performances for a set of two-scale models, in each of the image data sets. The unit of the spatial scale parameter is one pixel. The best classification performance occurs when the scale parameters were set at $m1/m2$ = 4/16 for the small molecule data, and $m1/m2$ = 6/20 for the protein data. Thus image features at spatial scales differing by a factor of around 3-4 are essential for

accurate classification in these image collections.

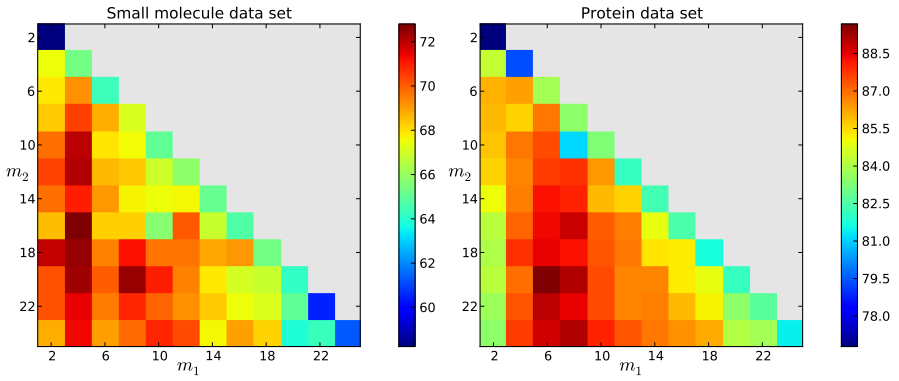


Figure 2.14: Classification rates based on two spatial scales in the small molecule data set (left) and in the protein data set (right).

As noted above, the overall prediction accuracy increases when using two spatial scales compared to using one spatial scale. Moreover, prediction accuracy at the individual category level either stays constant or improves when moving from one to two spatial scales. This holds for both the small molecule and protein data sets. We considered this further in the small molecule data by comparing the category-level prediction accuracies for three classifiers (m=4 alone, m=16 alone, and m=4,16). Some categories such as the mitosis category seem to rely on only one spatial scale (42% correct for m=4, 59% correct for m=16, and 60% correct for m=4,16). Another such example is the mitotracker red control category (81% correct for m=4, 70% correct for m=16, and 81% correct for m=4,16). In contrast, other categories seem to rely on a synergistic effect between the two scales: the plasma membrane category is correctly predicted in 45% of cases using either m=4 or m=16 alone, but this improves to 62% when both scales are used. Finally, some categories such as the autofluorescence category have nearly identical prediction rates for one or two spatial scales (88%, 87%, and 88% for m=4, m=16, and m=4,16).

2.4.2 Assessment of sensitivity to image feature parameters

After considering how the spatial scales influence classification rates, we then considered whether classification performance could be improved by optimizing the image processing or other image feature parameters. We did this by systematically varying a subset of the parameters while leaving the other parameters fixed.

We started the sensitivity analysis by image pre-processing: suppose x_{ij} is a pixel value, let $\alpha < \beta$ be two percentiles of the pixel intensity distribution, we considered the effect of the truncated linear scaling function

$$x_{ij} \to (\frac{x_{ij} - \alpha}{\beta - \alpha} \vee 0) \wedge 1$$

.

Choosing α from 0 to 0.2, and choosing β from 0.7 to 1, Figure 2.15 shows the result. We could see that the highest intensity present much more information then the lowest intensity, because the more fraction of intensity pixels removed from top, the lower the classification rate is, generally, for both data set, while in the same time, we could see that the variety of classification between different lower fraction removed is not as much large, and in several cases, the fraction removed most from bottom even behaves the best, which makes sense since we may removed much noise. In addition, we noticed that for both data sets, there is a jump between nothing removed from top intensity and a rather small upper fraction removed. The average bootstrap standard deviation for small molecule and protein data are around 1.4% and 0.98% respectively.

Similarly with the approach of exploring filter scale in Subsubsection 2.4.1, we intended to apply only one spatial aspect ratio γ (values from 1 to 10) in Gabor filter functions first, while keep the other parameters the same as baseline model. For the small molecule data set, the optimized performance was 0.6 percentage point drop

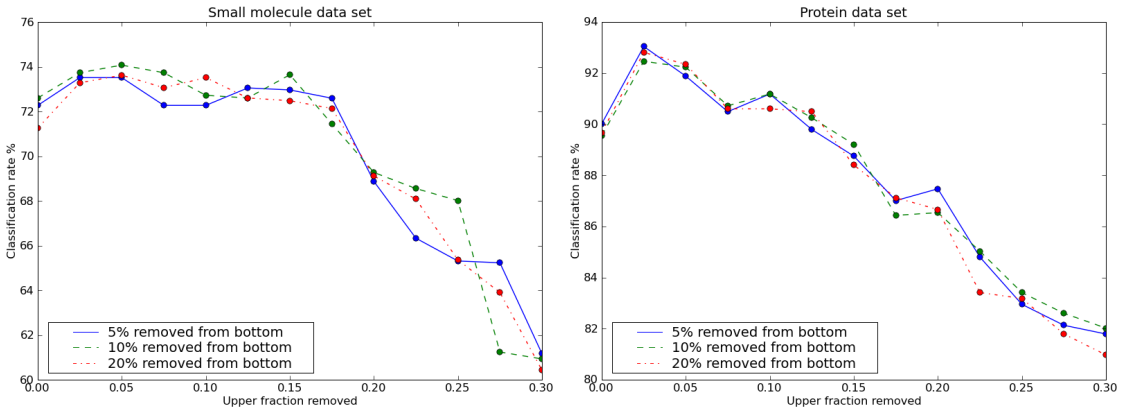


Figure 2.15: Classification rates under different truncated scaling ways. Left: small molecule data set; Right: protein data set.

in overall classification rate compared with baseline model, and the variation is not large within the classification rates by choosing one single spatial aspect ratio for both data sets. For the protein data set with original imaging magnification, the optimized classification rate was 6 percentage point drop in overall classification rate, which was not competitive to the baseline model, therefore, we considered choosing two ratios from 1 to 13, every other unit, and Figure 2.16 shows the result. The pair of (1,9) gave peak classification rates (74.82% and 90.02%) for the small molecule and the protein data set respectively, which was only 1.5 and 0.3 percentage point larger than by baseline model.

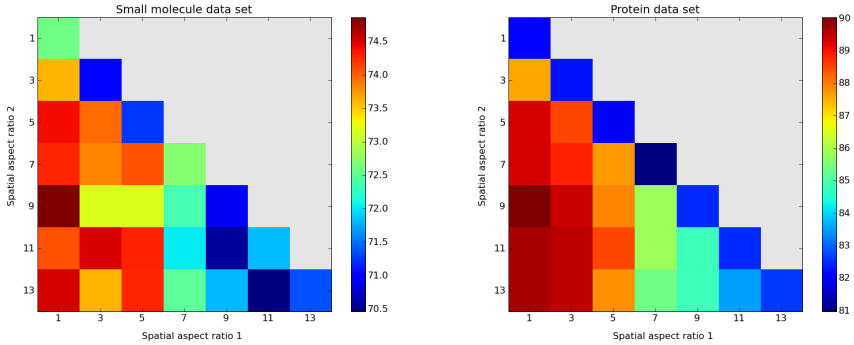


Figure 2.16: Classification rates using two spatial aspect ratios.Left: small molecule data set; Right: protein data set.

In the way of feature construction, we sliced the pixel pair into three sub-groups by using two quantiles of intensity, if we let Q_1 denote the lower quantile, and Q_2 stand for the upper quantile, then (Q_1, Q_2) pair is the 50th and 90th percentile of intensity in the baseline model. We also optimized the quantile thresholds, by ranging Q_1 from 10th percentile to 50th percentile, and ranging Q_2 from 60th percentile to 90th percentile, and the optimized results gave 1.2 percentage point increase in overall classification rate for the small protein data set, and 0.3 percentage point increase in overall classification rate for the protein data set.

As a conclusion, optimizing either the spatial aspect ratio, the intensity truncation parameters, or the quantile thresholds resulted in less than a 2 percentage point improvement in classification performance in either data set. These improvements are comparable to the standard errors in the classification rates as quantified by the bootstrap procedure.

In order to better understand how each feature differs in difference categories, we calculated two-sample z test statistics for each 108 features in the small molecule data set whole cell region and in the protein data set (Figure 2.4.2). Through these two graphs, for certain groups, like 'autofluorescent' in the small molecule data set or 'nucleus' in the protein data set, they have obvious difference with other groups under a few features. This evidence provides somewhat support for our previous hypothesis that these conditional texture-based features could reveal certain localization information.

2.4.3 Haralick and Zernike features

Haralick texture features [24] contain information about the spatial distribution of gray tones, the structural organization of surfaces and relationship among their neighborhoods. Zernike features [31] are a set of shape features based on Zernike polynomials using the coefficients as features. The CellProfiler software [42] was

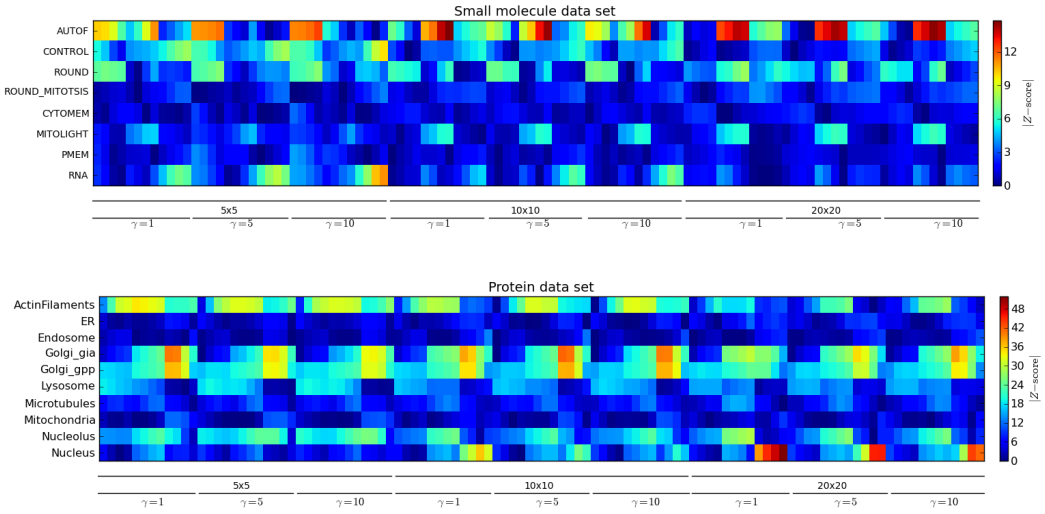


Figure 2.17: Z-scores for each feature in the two data sets. Top: the small molecule data set. Bottom: the protein data set.

used to calculate sets of 14 Haralick and 30 Zernike features (at levels 0 to 9). These features were then used with leave-one-out LDA for image classification, then the classification rates and standard deviations of bootstrap sampling were 44.40%, 77.11% and 2.890%, 4.442% for small molecule and protein data set respectively.

2.4.4 Regularization of discriminant analysis

Two elementary and popular methods of classification are linear discriminant analysis(LDA) and quadratic discriminant analysis(QDA). LDA was used to find the optimum projection directions to make the within-class scatter denser and between class more scattered, and diagonal discriminant analysis works on the diagonal entries on the high-dimensional covariance matrix. If we let $\hat{\Sigma}_b$ denote the between-group variance matrix of the training data, and $\hat{\Sigma}_w$ denote the within-group variance matrix of the training data, and $\hat{\Sigma}_{w,D} = diag(\hat{\Sigma}_w)$, then the criteria of LDA is to find $\arg\max_{v=1} \frac{v'\hat{\Sigma}_b v}{v'\hat{\Sigma}_w v}$, and method DLDA is to find $\arg\max_{v=1} \frac{v'\hat{\Sigma}_b v}{v'\hat{\Sigma}_{w,D} v}$. If we let a parameter $\lambda \in [0,1]$, $\hat{\Sigma}_{w,\lambda} = \lambda\hat{\Sigma}_w + (1-\lambda)\hat{\Sigma}_{w,D}$, then the regularization between LDA and DLDA would become to find $\arg\max_{v=1} \frac{v'\hat{\Sigma}_b v}{v'\hat{\Sigma}_{w,\lambda} v}$. Similarly, we could also carry on the

regularization between LDA and identity matrix, between QDA and diagonal QDA (DQDA) as well. The protein images used here were down-sampled by 2*2.

The criterion to judge the performance of one method is the true error rate, however, we are unable to obtain the true error rate, we could only use crossvalidated error rate to estimate it. Therefore, all the decision and conclusion are based on the estimated error in the following document, and the error rates in the following document all refer to the leave-one-out errors if not specified. Within the family of regularized classification rules spanning from diagonal LDA or identity matrix to standard LDA, the regularization results can be shown on Table 2.4 and Table 2.5, based on these, the standard LDA method gives almost the best performance for both the small molecule data set and the protein data set, this is likely attributable to the moderately strong correlations among the features that result from the overlapping shapes and scales of the various filters that were applied to the images.

2.5 Visualization analysis

To complement the classification analysis, we also considered the performance of our features for use in dimension reduction and data visualization. In principle, the protein data set should be able to serve as an atlas of subcellular structures onto which the small molecule staining patterns can be mapped. Figure 2.18(a) shows the distribution of protein images on the dominant two principal components (PC's) of the protein data, using our feature set. The nuclear and actin cytoskeleton localization patterns are clearly separable from each other, and from the other localization categories. Within the other eight categories, some additional segregation of classes was observable, but for this analysis we will not focus on these differences.

A key question of interest was the relationship between the small molecule and GFP-tagged protein image sets. To address this, we overlaid the small molecule images on the PC's defined by the protein data. Figure 2.18(b) shows that the styryl

Table 2.4: Regularization for LDA, each entry is a percentage.

		λ*LDA+(1-λ)*Regularization										
Data set	Regularization	λ=0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Small molecule	DLDA	52.31	62.34	66.52	67.42	68.55	70.25	71.48	72.38	73.51	75.54	73.06
	Identity	45.10	45.88	46.67	47.69	47.91	48.14	47.91	48.82	50.17	52.20	73.06
Protein	DLDA	71.11	74.94	76.91	78.07	79.93	80.39	82.95	83.41	83.76	85.96	88.52
	Identity	63.57	65.20	65.66	66.94	68.44	68.79	69.95	71.23	73.66	76.57	88.52

Table 2.5: Regularization between QDA and diagonal QDA, each entry is a percentage.

	λ*QDA+(1-λ)*DQDA										
Data set	λ=0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Small molecule	52.20	59.08	60.88	61.44	61.33	61.22	60.99	60.77	60.09	58.62	10.26
Protein	73.20	82.48	84.45	85.61	86.19	86.54	87.24	88.05	87.94	88.63	1.51

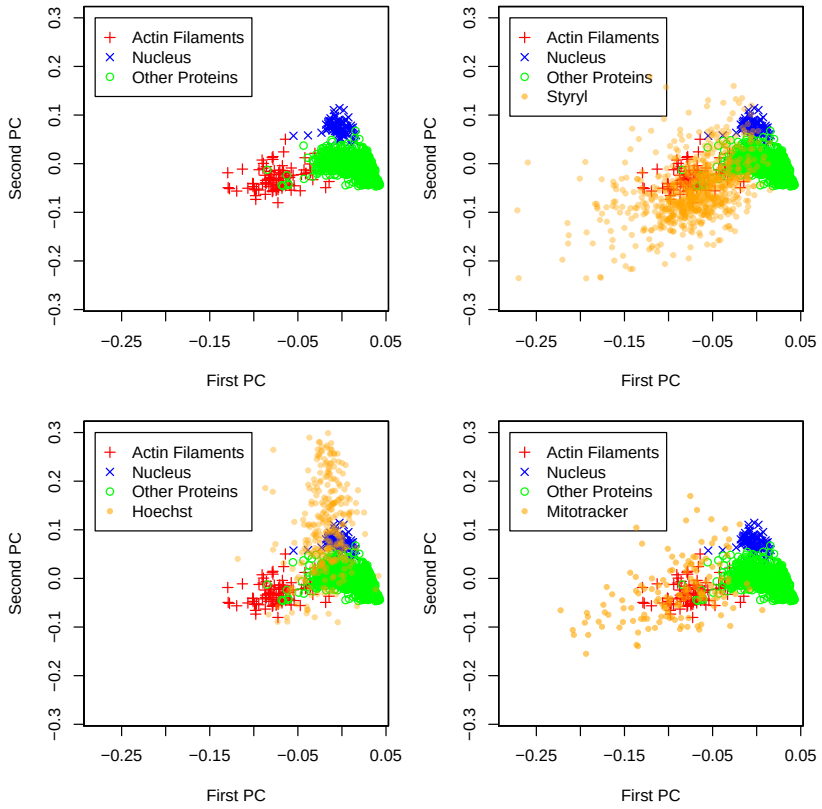


Figure 2.18: Principal Components Analysis of the protein data, superimposed with various subsets of the small molecule data: a) the protein images alone, b) the protein images with styryl images superimposed, c) the protein images with Hoechst images superimposed, d) the protein images with mitotracker red/green images superimposed.

molecules are strongly segregated from the organelle-specific GFP reference markers, but show considerable overlap with GFP-labeled actin cytoskeleton reference patterns. As a positive control, we also confirmed that the Hoechst-labeled cell nuclei in the small molecule imaging experiment fall in the same region of image feature space as the GFP-tagged proteins localizing to cell nuclei (Figure 2.18(c)). As separate controls, cells stained with the mitotracker red imaging probe broadly occupy the same region of image feature space as the styryls.

We performed an additional numerical experiment to exclude the possibility that the relationship between the small molecule and protein data sets was an artifact

of the lower signal-to-noise level in the small molecule data set. We added multiplicative log-normally distributed noise to the protein images (i.e. x was transformed to $exp(log(x) + e)$, where e is normally distributed with mean zero and variance v. We then repeated the feature construction, and calculated the Euclidean distance in image feature space from the centroid of the styryl images to the centroid of each localization class of the protein images, and to the overall centroid of the protein images. These distances were found to be a strictly increasing function of v. We note that due to the non-linearities in the feature construction, this was not a foregone conclusion, and indeed we were able to construct synthetic data sets in which differing noise levels strongly influenced the relative positions of the images from different data sets. The observation that the styryl localization patterns become less similar rather than more similar to any part of the protein data set suggests that the actin-like localization pattern of the styryl probes is not an artifact of the differing noise levels inherent in these data.

CHAPTER III

Measurement Errors and Artifacts in High Content Imaging

High content images are subject to several major forms of artifacts and noise. The images we focus on here are obtained using fluorescent reporters (either fluorescent small molecule probes, or endogenous fluorescent reporters such as green fluorescent protein, or GFP). Ideally, the fluorescence intensity is directly proportional to the concentration of the underlying molecule of interest. For a number of reasons, this linear proportionality, or even a monotonic approximation to it, may not hold in real experimental data. Deviations from the ideal proportionality can often be attributed to specific sources of errors and artifacts in the imaging process. In this chapter, we consider the implications of measurement errors and artifacts on our ability to draw scientifically meaningful conclusions from high content imaging data.

To illustrate, Figure 3.1 shows several high content images that are affected with various types of artifacts. The left panel shows an image that was acquired out of focus, the center panel shows an image that was acquired with high image noise, and the right panel shows an image that was was subject to a high level of saturation.

In dealing with image noise and artifacts, we cannot in general hope to restore the images to their ideal "true" state after they have been corrupted in some way. However, we can take advantage of the fact that in HCS studies we have a large

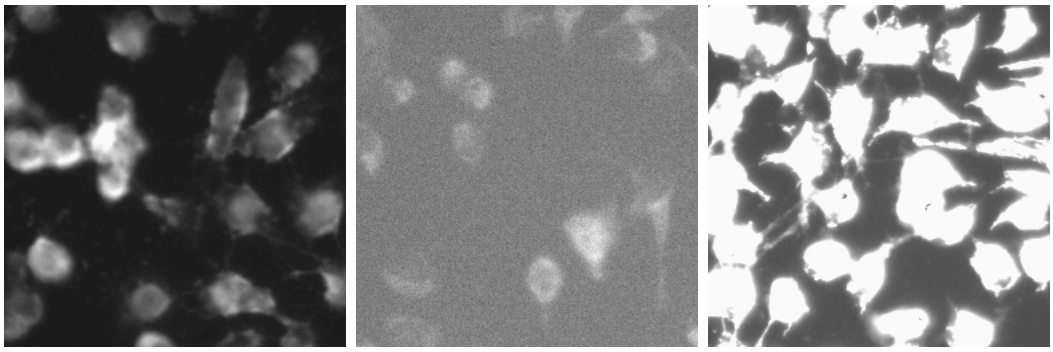


Figure 3.1: Examples of high content images that are affected with various types of artifacts. From left to right, the types of artifacts are: out of focus, image noise and a high level of saturation.

collection of images, which can be presumed to be subject to similar forms of artifacts. Our ultimate interest is in some low dimensional summary statistic of this image collection, rather than in the images themselves. As we will explore further below, the consequences of various forms of imaging artifacts can usually be viewed as introducing bias into the image summaries that are used in the downstream scientific analysis. Implications for variance are often secondary. Thus, we can view these imaging errors and artifacts using a statistical framework for analyzing bias, building on other successful instances of bias reduction in the statistics literature.

As we will see, it is often possible to assess the direction, and sometimes the magnitude of the bias on downstream summaries resulting from imaging errors and artifacts. Based on an understanding of the bias, it then becomes possible to consider how it could be handled. For example, suppose we are doing group-wise comparisons. If all groups are affected in similar ways by an imaging error or artifact, the bias will have minimal impact on our ultimate conclusions, and thus can generally be safely ignored. But if the bias differs among the groups, it can either exaggerate or attenuate the inter-group differences. Thus, we place a special emphasis on identifying situations where such differential bias is present.

The study of measurement errors is highly developed in the regression analysis set-

ting, where measurement errors in the independent variables play a very different role than the additive errors in the response that were classically the sole source of variation in regression models. A number of techniques including regression calibration, and the "Simulation Extrapolation" (SIMEX) method, were developed specifically to address the bias that results from covariant measurement errors in regression analysis [25] [43] [17]. While these methods cannot be directly applied to the very distinct setting of analyzing image collections, we will see that with some adaptations, tools from regression analysis can be used in this context.

Since our ultimate aim is to assess specific scientific hypotheses using the images, we should carefully consider the impact of any approach to bias reduction on our ability to meet our scientific aims. In particular, it is likely that any efforts to reduce the impact of a particular imaging error will increase the estimation variance. In cases where the bias may not have a large impact on our findings, it is possible that we will not benefit from efforts to reduce bias. However as we shall see, there is good reason to believe that bias reduction is beneficial in at least some settings.

A final comment about addressing measurement errors and biases in an image analysis setting is that our goals are broader than producing specific methods for bias elimination or bias reduction. In addition to these goals, we can also consider a more exploratory sensitivity-analysis perspective on bias issues. The reason for doing this is that even if the substantial technical challenge of reducing or eliminating the bias were possible, to actually apply such a bias adjustment in practice would require an accurate mathematical description of the exact error process that is actually present in a particular image collection. In current high content screening, this may not be a realistic expectation. For example, as discussed below, it is reasonable to expect that a certain amount of blurring is present in the images. But the exact amount of blurring may be difficult to quantify. Taking a sensitivity analysis perspective, we can assess the implications of having various plausible levels of blurring in the images.

Knowing the extent to which these various levels of blurring impact downstream analysis gives us a good sense of how much additional uncertainty due to imaging errors may be present in our downstream results, even if we are unable to correct for it.

This chapter is organized as follows. Section 3.1 considers three major forms of imaging errors and artifacts that we will use to illustrate our analytic approaches. This section also demonstrates how we will assess the impact of imaging errors on downstream analysis, and illustrates the approach using multiclass distributions of vectors. Sections 3.2 through 3.4 consider in detail how specific characteristics of the image population are affected by imaging errors. Each of these sections considers a distinct descriptive characteristic. Synthetic images and real data are used in each section to illustrate the impact of imaging errors, and to explore how this impact can be recovered in our analysis.

3.1 Sources of errors and artifacts

High-content imaging, like many forms of high-throughput screening, can be viewed as an experimental trade-off in which the experimenter forgoes a high level of data quality in order to obtain a large volume of data. The usual premise of this tradeoff is that in a screening context, there is a relatively higher tolerance for spurious positive results, as it is relatively inexpensive to validate and confirm all positive results before proceeding with traditional low-throughput analysis. In the imaging context, this means, for example, that resolution, magnification, and exposure times are limited, to enable many wells on a micro-titre plate to be imaged in a reasonable period of time. This directly affects the image contrast and the signal-to-noise ratio, since a lower number of detections from the true biological sources is expected, and hence the biological signal of interest is relatively smaller in comparison to the noise and artifacts in proportion to detections that result from errors or artifacts. In one

of our experimental data sets, we are focusing on the localization behavior of small molecules, so as to understand how the structure of a chemical probe influences how it localizes in a cell. Since most compounds of scientific interest are at best weakly fluorescent, we are dealing with low-contrast, low-signal images out of necessity.

It would be possible to identify many plausible sources of errors and artifacts in high content images. Ideally, we would like a general methodology for accommodating any source of errors or artifacts that can be statistically characterized. It is unclear whether this is a realistic aim, but it does seem possible to develop frameworks that are applicable to several sorts of errors. To demonstrate the potential of doing this, we will focus on three specific imaging errors that are highly distinctive in statistical terms, and that are undoubtedly present in high content imaging data. We refer to these artifacts as "saturation", "blurring", and "noise", and describe them in more detail in the following sections.

Saturation

The fluorescence signal in a high content imaging study is detected by a digital CCD, which becomes saturated once the signal reaches a maximum value (for example, either 4095 for a 12 bit detector, or 65535 for a 16 bit detector). The pixel intensities take on the ceiling value exactly when the true signal exceeds the ceiling. This clearly violates the ideal proportionality between detected fluorescence and probe concentration. The visual manifestation of this artifact is the appearance of flat white patches in the images. While the functional form of saturation is a simple step function, we can anticipate that due to its non-linear nature it may have complicated and difficult to characterize implications for our downstream results. One favorable aspect of this type of artifact is that we will in practice know its exact functional form, since the truncation occurs exactly at the maximum capacity of the detector, which is a known quantity.

Blurring

A CCD is a grid of discrete detectors, and the accompanying optics are designed to gather fluorescent light from specific physical locations on the imaged surface and focus it onto a corresponding detector. Due to imperfections in the optics, and other forms of scattering and diffusion, signals from a specific location in the image target will not always be detected at the appropriate point on the CCD. Much of this scattering will be a small perturbation from the ideal, and will result in signals will be detected close to the appropriate site. This artifact can therefore be viewed mathematically as a form of convolution, or blurring, in which the observed image is obtained by convolving the ideal image with a filter whose shape and width reflects all sources of scattering and signal diffusion. Visually, this artifact appears as the typical form of image blurring in which edges become less distinct, and the contrast between foreground and background objects is lowered. The linearity of this particular artifact might be expected to produce a more continuous or analytically tractable impact on downstream results, compared to the saturation effects discussed above. In practice, we may have a rough idea of the extent of blurring, for example, through calibration studies in which a phantom object is imaged and the acquired image is compared to the ideal image. However such phantom experiments may not reflect the reality of image acquisition with live cells, so to some degree the level of blurring in actual data will be unknown.

Additive Noise

There are various sources of fluorescence in live cell studies besides the sources of interest, including cellular autofluorescence and emission from the non-biological material holding the sample. In addition, there is likely to be some form of electrical "shot noise" in the instrument itself. These sources of error can be viewed as statistical noise that is approximately additive to the signal of interest at the level of individual pixels. This source of error is thus most analogous to the traditional statis-

tical notion of additive measurement error. Some important aspects when considering the effects of this additive noise are the marginal distribution of the noise (i.e. the distribution within each pixel), the dependence of the noise in space (between pixels), and, if relevant, the temporal dependence. Some previous work in this area has focused on detailed physical models for imaging noise that use Poisson-like distributions to capture individual signal events (e.g. photon detections). An alternative approach is to view the data as being aggregated to a point where a continuous distribution such as a Gaussian distribution can be used to describe the noise at each pixel. The visual impact of additive white noise is a degradation of sharp edges, and a general weakening of contrast between distinct objects. Non-white additive noise can introduce artificial objects into cells. The additivity of this type of artifact is expected to make it somewhat more tractable to accommodate in downstream statistical analysis. In practice, we can estimate the noise level by considering the acquired signal in background pixels. While this gives a rough estimate of the noise level, it may not be constant in the image, in particular since some sources of noise may arise from the cells or be impacted by the presence of a cell. Thus there will be some uncertainty in practice as to the level of additive noise in a given image collection.

An important issue in considering these sources of errors and artifacts is that they are all taken to operate directly at the pixel level in the images. In practice, the first step in most instances of scientific image analysis is to convert the images into relevant features. These features are nearly always nonlinear functions of the pixel intensities, since linear functions would generally not be invariant to the positions of the cells in the well, which are not controlled experimentally. Thus we are faced with a form of error propagation. Even when a particular source of pixel-level error or artifact is well-behaved, such as being linear, monotonic, or additive, its impact on the features may be much more complex. For example, we may be interested in a simple summary statistic such as the mean of a particular image feature within

a class of images (e.g. within all images in which the actin cytoskeleton is stained). Although the imaging noise is additive at the pixel level, and hence would have a vanishing impact on any linear function of the pixel intensities as the sample size increases, this vanishing behavior is not expected to hold at the feature level.

In practice, we anticipate that these three sources of image artifacts are simultaneously present, along with numerous other known and unknown sources of artifacts. Moreover, there may in some sense be interactions between various artifact-generating processes. Specifically, it is natural to ask in what sequence these artifacts are introduced. It seems most logical for saturation to be the final operation in this sequence. But it is less clear whether blurring follows additive noise, or vice versa. We do not consider these sequencing effects further, but note that this is an important area for future work.

3.1.1 Simulation-based approaches to bias analysis

In principle, the impact of most forms of imaging errors and artifacts on downstream analysis could be assessed using either mathematical analysis or simulation. For example, in the well-explored area of measurement errors in regression analysis, the regression calibration approach relies more on mathematical analysis, while the "simulation-extrapolation" (SIMEX) approach relies more on simulation [11] [23] [25] [43]. A generally applicable approach to handling imaging errors that is primarily based on mathematical analysis seems difficult to achieve. We thus primarily focus on simulation based approaches here.

A basic simulation based approach to understanding the impacts of imaging artifacts on downstream analysis begins by considering a true image I, and an artifact that can be represented as a function $T_\theta(\cdot)$, so that $T_\theta(I)$ is the image with the artifact present. The function T_θ may be either deterministic, as in blurring, or random, as in the addition of additive noise to the pixels. The parameter θ can be viewed

as reflecting the strength of the artifact, so that $T_0(I) = I$, and greater values of θ represent a greater deviation between I and $T_\theta(I)$. In some cases the value of θ will be known, either exactly or through an accurate approximation. More commonly, we will have only a rough idea about the value of θ. For example, we may know that θ lies in a certain interval, or we may have either an upper or lower bound for the value of θ (but not both).

The goal of our analysis is to consider the effect of the artifact represented by T_θ on a summary statistic $F = F(I_1, \ldots, I_m)$, where the I_j are a collection of images. Specifically, if $J_j = T_\theta(I_j)$ is the j^{th} observed image, a naive analysis results in the statistic $\tilde{F} = F(J_1, \ldots, J_m)$. As noted above, the discrepancy between $\tilde{F}$ and F might most effectively be viewed in terms of its mean value, interpreted as a bias, even in the setting where T_θ is a random function. In principle, impacts on the variance would also be an interesting object of study, but we do not consider this further here. This perspective is consistent with the usual practice for errors in variables problems in regression analysis, where the focus is much more on bias than on variance.

In practice, we observe the images $J_j = T_\theta(I_j)$, not the original images I_j. For a hypothetical value of θ, we can then form $\tilde{J}_j = T_\theta(J_j) = T_\theta(T_\theta(I_j))$. We can then consider the difference between $F(\{\tilde{J}_j\})$ and $F(\{J_j\})$, which is observable, as a proxy for the difference between $F(\{J_j\})$ and $F(\{I_j\})$, which is not observable. This may give us some information about the direction and magnitude of the bias resulting from application of $T_\theta(\cdot)$ to the images.

Initially, we can consider this as a form of sensitivity analysis for learning about the biasing effect of T_θ. More ambitiously, we can use the quantity

$$B_\theta = F_\theta(\{\tilde{J}_j\}) - F_\theta(\{J_j\})$$

as an estimate of the bias, and subtract it from the naive estimate to produce the bias-reduced quantity

$$F_\theta(\{J_j\}) - B_\theta = 2F_\theta(\{J_j\}) - F_\theta(\{\tilde{J}_j\}).$$

This is precisely the idea behind the SIMEX procedure for reducing the impact of covariant measurement errors in regression analysis.

Successfully applying a SIMEX procedure to reduce bias requires two conditions. First, a reasonable idea about the value of θ must be available. Second, we are implicitly assuming a linearity in that the incremental impact of successive applications of the transformation T_θ is constant. More commonly, the reality is that each successive application of T_θ has a diminishing effect. In other words, the function B_θ is convex in θ. As a result, this simple bias adjustment typically undercorrects on average for the bias. In some cases it might be possible to estimate the curvature in B_θ so as to more completely accommodate the bias. While this has been effectively done in the regression setting, the image analysis setting presents a number of other difficulties, for example, the fact that we may have only limited information about the range of plausible values for θ. As a result, we do not pursue efforts to precisely remove the bias here, but rather focus on a more exploratory style of sensitivity analysis in which the goal is to estimate the direction, and to a rough degree, the magnitude of the bias. This should allow us to assess whether bias is a major factor in a given inferential setting, even when we cannot remove the bias. This in turn can guide the overall direction of scientific investigation, in that questions that are especially sensitive to bias might be set aside if other questions of interest do not have this liability.

3.1.2 The downstream analysis

The impact of imaging errors is a consequence both of the properties of the error-generating process, and of the specific downstream analysis that is to be performed. For instance, additive errors with zero mean will have minimal impact on a downstream procedure that primarily focuses on linear functions of the pixel intensities,

but will have a much bigger impact on a downstream procedure that focuses on dispersion. It is less clear how to attain an intuitive sense of the impacts of blurring and thresholding on various types of downstream analysis.

Measurement errors have been extensively considered in contexts where the goal is point estimation, and formal inference such as hypothesis tests and confidence intervals. A classic example is the attenuation of Pearson's correlation coefficient when the measurements are affected by additive noise. The resulting bias also implies that hypothesis tests have reduced power, and confidence intervals have reduced coverage probabilities.

Here we take a different aim, and focus on the impacts of measurement errors on multivariate, graphical, and descriptive statistics. As a very simple introductory example, we can consider the distance between the centroids in feature space of various classes. Considering $K > 2$ classes, there are $\binom{K}{2}$ pairwise distances, and the relationships among these pairwise distances convey which pairs of classes are relatively more similar to each other compared to other pairs. Figure 3.2 shows several examples of pairs of classes with different inter-centroid distances, from short, moderate to relatively long distances, accordingly. If imaging errors affect these pairwise distances differentially, it is possible that our perception about which classes are most similar to each other will be systematically wrong. We will consider whether simulation-based sensitivity analysis can be effectively used to assess whether this is happening in a given data setting.

A more complex situation is the widely-used technique of Principle Components Analysis (PCA). PCA can be effectively used to summarize the relationships between different classes of images, as we did in Chapter 2 to assess how the small molecule localizations relate to the GFP-labeled protein localizations. In this setting, interpreting the results of PCA largely focuses on the relative distributions of the points in different groups. Attention will naturally focus on the the pairwise distances be-

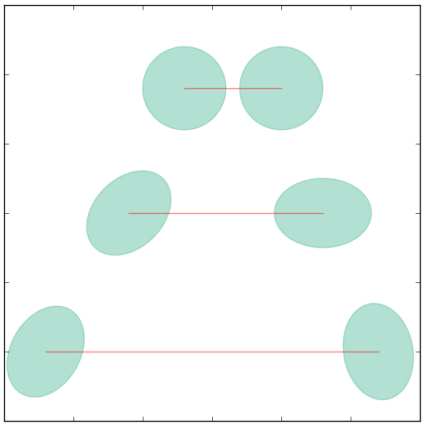


Figure 3.2: Distances between class centers in the PCA projections.

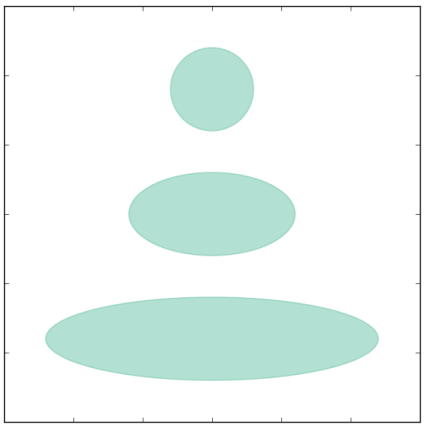


Figure 3.3: Eccentricity of the distribution of one class in the PCA projection.

tween class-specific centroids, as discussed above. Beyond this, we can consider other aspects of the distribution of images in PCA space, for example, the degree of eccentricity in the class-specific distributions, as depicted in Figure 3.3, from circle, ellipse to elongated ellipse, accordingly; and the angles between the dominant axes of variability for different classes, as depicted in Figure 3.4, from parallel, intersect to perpendicular, respectively.

We began our study of this phenomenon by considering an artificial data setting in which the data are vectors rather than images. To be specific, we constructed three classes, each class is a collection of independent multivariate normal random vectors,

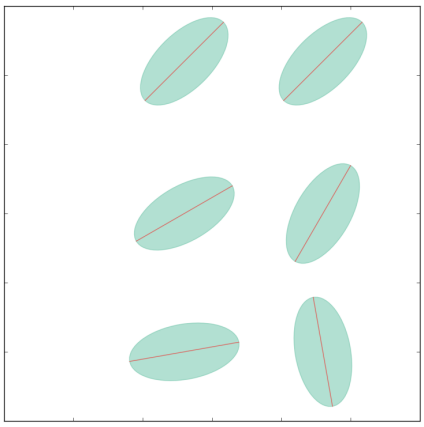


Figure 3.4: Changes in the angles of dominant directions for two classes in the PCA projection.

for each vector, the length is 100, and each class includes 200 independent random vectors, thus each class can be treated as an artificial 'image' of 200×100. When making a simple scatter plot projected into two-dimensional principle component space, they appear as in Figure 3.5, where the numerical label associated with each ellipse represents the number of class. In the two-dimensional projection, these three classes have well-separated centers. The degrees of eccentricity of classes 2 and 3 are similar, but are dramatically different from class 1. The dominant axes of variability for classes 2 and 3 are approximately parallel, but are very different from class 1.

We first considered calculating the pairwise centroid distances for the original unperturbed data as in Figure 3.5, then calculating the inter-centroid distances after artifacts introduced by blurring alone, adding noise alone, and thresholding alone (from left to right). Figure 3.6 illustrates how the distances change under different situations (for Figure 3.5 and Figure 3.6, colors represent the classes). We can observe that the pairwise centroid distances are largely depend on the degree of blurring, be more specific, they decrease monotonely as amount of blurring increase, and within certain range, they vary linearly. The amount of additive noise does not influence the pairwise centroid distances much, as the distances stay approximately constant relate

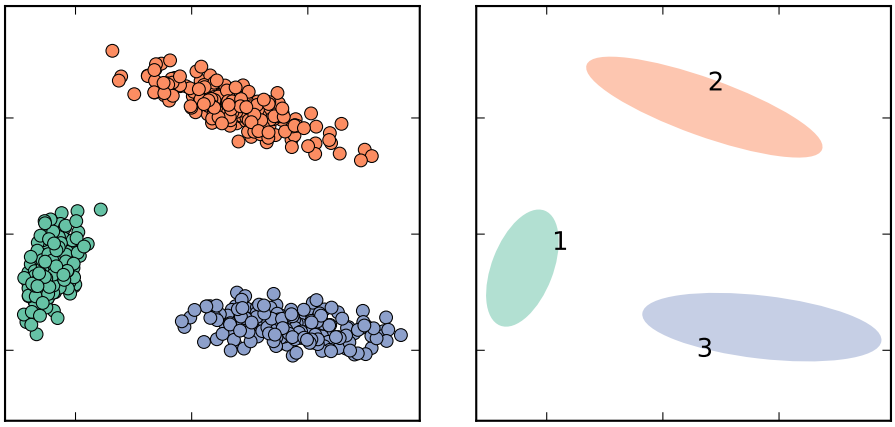


Figure 3.5: Three classes of simulated multivariate normal random vectors, projected into two-dimensional PC space.

to the amount of additive noise. This is expected since the pairwise centroid distances are directory calculated based on 'pixel' values, and when each 'pixel' values is added with a random normal variable with mean 0, the average number added to the center of class is approximated 0, thus the centroid distance would not change a lot. When considering truncation of random vectors (correspond to saturation in real images), within the range of the right end which is associated with not truncated much, we would considered that the pairwise centroid distances stay almost constant. But if the strength of truncation turns to be moderate, the distances drop dramatically. All these observations from Figure 3.6 supports the use of linear approximations when applying the SIMEX procedure, at least in this simple setting.

Next we considered the eccentricity of each class and the angles between dominate axes of variation for pairwise classes. We plotted the results in Figure 3.7, Figure 3.8 and Figure 3.9 (the original unperturbed data in these three figures is the same as data in Figure 3.5). These figures shows how each class of random vectors changes when it has been affected by more and more blurring, additive noise, or truncation respectively. For each figure, the subplots from left to right represent class 1, 2 and 3. Within each subplot, the numerical label presents the strength of perturbation. Five

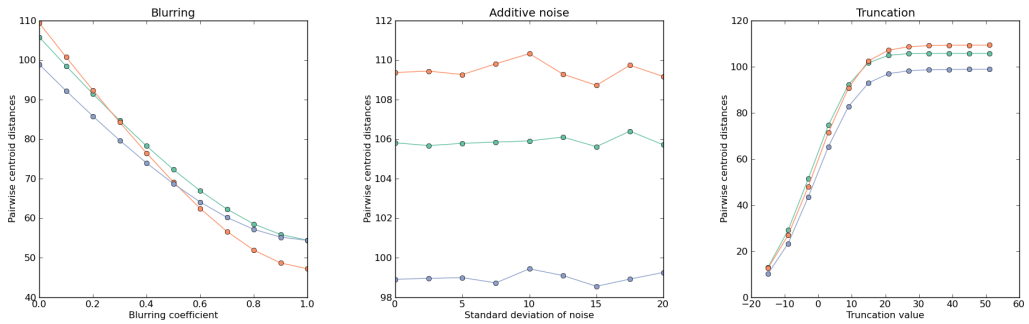


Figure 3.6: The distances between centroids of three classes of the random vectors with artifacts introduced, from left to right: blurring, additive noise and truncation.

different strength have been plotted. Strength 1 represents the original unperturbed data, and the larger value is subject to more amount of artifacts. For these five strength, the set of associated parameters is a subset of the parameter values appear in Figure 3.6, Figure 3.10 and Figure 3.11. From Figure 3.7, it is noticeable that the degree of eccentricity changes obvious when blurring is introduced, along with the angles between dominant axes of variation for different classes. From Figure 3.8, the degrees of eccentricity does not have a dramatic change when random noise is added, the angles of dominate axes of variation does not change much either when standard deviation of additive noise is not very large, but when the according standard deviation continues to be increase, the angles of dominant axes of variation appear obvious change as we noticed for strength 5. From Figure 3.9, when the classes are affected by moderate strength of truncation, the degree of eccentricity and angles of dominant axes of variation do not change a lot, but when the strength of truncation is very strong, the classes in the two-dimension projection turn to be more and more elongated. This is expected since when a large amount of values in a class has been truncated, the class turns to be more and more like a one-dimension class, and the variation becomes much smaller, thus we observe three much less expended (as variation is much smaller) but much more elongated ellipses for each of three

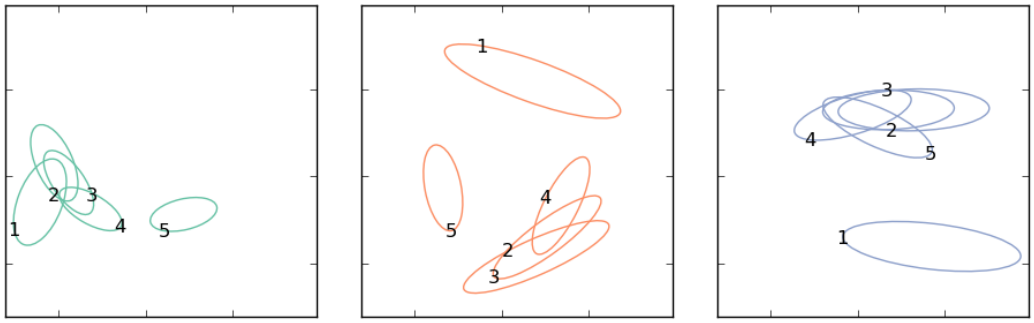


Figure 3.7: How the three classes change when they are introduced by more and more blurring.

classes at the strength of 5.

Be specific, we refer to Figure 3.10 and Figure 3.11. These figures display numeric changes of the degree of eccentricity and relative angles of dominant axes of variation for pairwise classes. From Figure 3.10, for blurring, the linearity is not as regular as in Figure 3.6 which is a monotonely decreasing linear trend, but we if we consider sub-intervals (such as below and above 0.45), linearity still holds locally. For additive noise, the overall trend is that the projected classes turn to be more and more circular, with local linearity appearing. For truncation, the degree of eccentricity decrease slightly and linearly when about half amount of the values in the classes are truncated (we expect this because the mean value of each class is 0 and when truncation value is not less then 0, we observe such fact), when more than half amount of values are truncated (truncation values smaller than 0), the projected classes become more and more elongated ellipses, as we stated in the above paragraph. From Figure 3.11, the relative angle of dominant axes of variation for pairwise classes change linearly when the classes are more and more blurring, they also change approximately linearly when more and more noise is added. There is no overall regular pattern when the classes are truncated, but when the amount of truncation is not very large (truncation values greater 20), a local linear trend appears.

In this section, we simulated three classes of multivariate normal random vectors

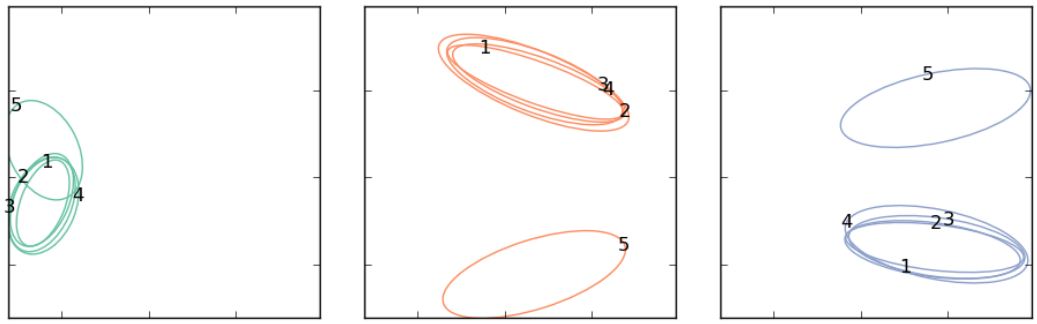


Figure 3.8: How the three classes change when they are introduced by more and more additive noise.

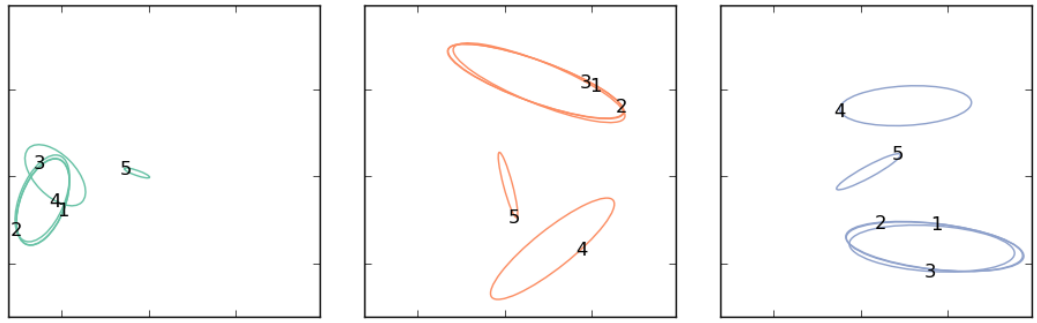


Figure 3.9: How the three classes change when they are introduced by more and more saturation.

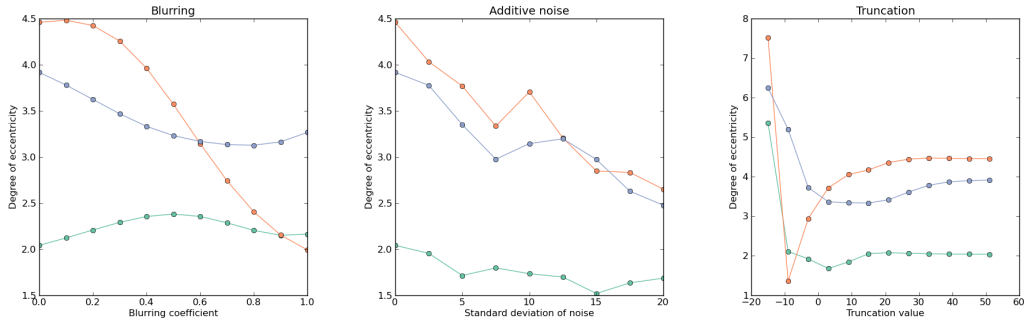


Figure 3.10: The degree of eccentricity of three classes of the random vectors with artifacts introduced, from left to right: blurring, additive noise and truncation.

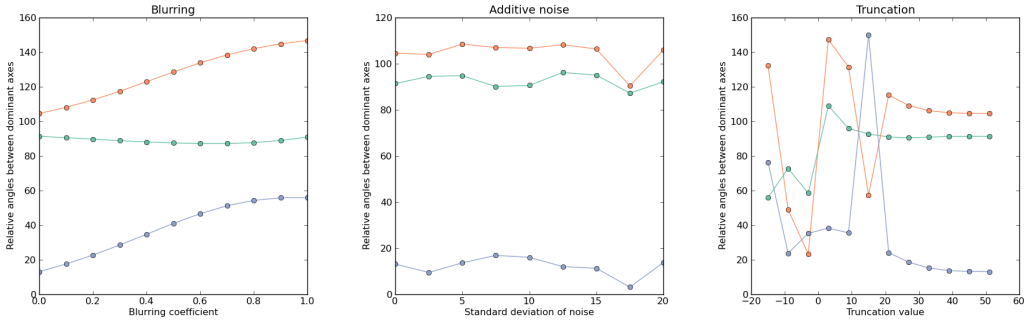


Figure 3.11: Relative angles of dominate axes of variation for pairwise classes of the random vectors with artifacts introduced, from left to right: blurring, additive noise and truncation.

and considered three statistics, the pairwise inter-centroid distance, the degree of eccentricity and the relative angles of dominant axes of variation for pairwise classes, we analyzed the impacts of measurement errors for these three statistics from three different sources of errors, blurring, additive noise and truncation. Local linear trends have been observed due to that we have a very simple date setting.

In the following sections, , similar analysis will be applied to the protein data set and artificial images. In these sections, the statistics are calculated based on image features instead of raw 'pixel' values in this section, as the features are not a linear function of pixel values, we can expect that the impacts of measurement errors would become much more complicated. For Section 3.2, Section 3.3 and Section 3.4, the analysis of pairwise centroid distances, the degree of eccentricity and the relative angles of dominant axes of variation will be given, respectively. Within each section, the analysis of statistics under three different measurement errors will be given in a order of blurring, additive noise and saturation, starting with the protein data set, followed by the artificial images.

3.2 Measurement errors for pairwise centroid distances

Blurring

Suppose we use I to denote the original observed images, h as the kernel function, then $g := I * h$ is the blurred version of image I by convolving kernel h. We applied Gaussian function

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

as the blurred filter function which is space invariant and the model generates a linear relationship between I and g, where x and y are the distances from the origin in the horizontal and the vertical axis respectively, and σ is the standard deviation of the Gaussian distribution.

Gaussian blur is a low pass filter, which removes fine image details but passes low-frequency signals, by generating the new pixel value as the result of weighted average of that pixel's neighborhood. The weight is depend on the distance from the original pixel. Specifically, the proportion of the second largest weight to the largest weight is $e^{-\frac{1}{\sigma^2}}$. And when σ goes larger, we can assume that the convolved images through Gaussian filter would become more smooth.

An image-based setting is considered further below. Due to ten classes existing for the protein data set, therefore, a total of 45 possible pairs of distances existing. We selected three classes (ER, Microtubules and Mitochondria) from the protein data set as an example. Figure 3.12 gives an example of cells for each of the three classes, from left to right: ER, Microtubules and Mitochondria; from top to bottom: original images, blurred images with $\sigma = 1.2$ corresponding to $e^{-\frac{1}{\sigma^2}} \approx 0.5$, blurred images with $\sigma = 12$ corresponding to $e^{-\frac{1}{\sigma^2}} \approx 0.99$. It is noticeable that when $\sigma = 1.2$, the overall pattern of images were still recognizable, but the fine structures were blurred; when $\sigma = 12$, it generated almost white-black images without recognizable structure.

The pairwise centroid distances are plotted at the left panel of Figure 3.13, the

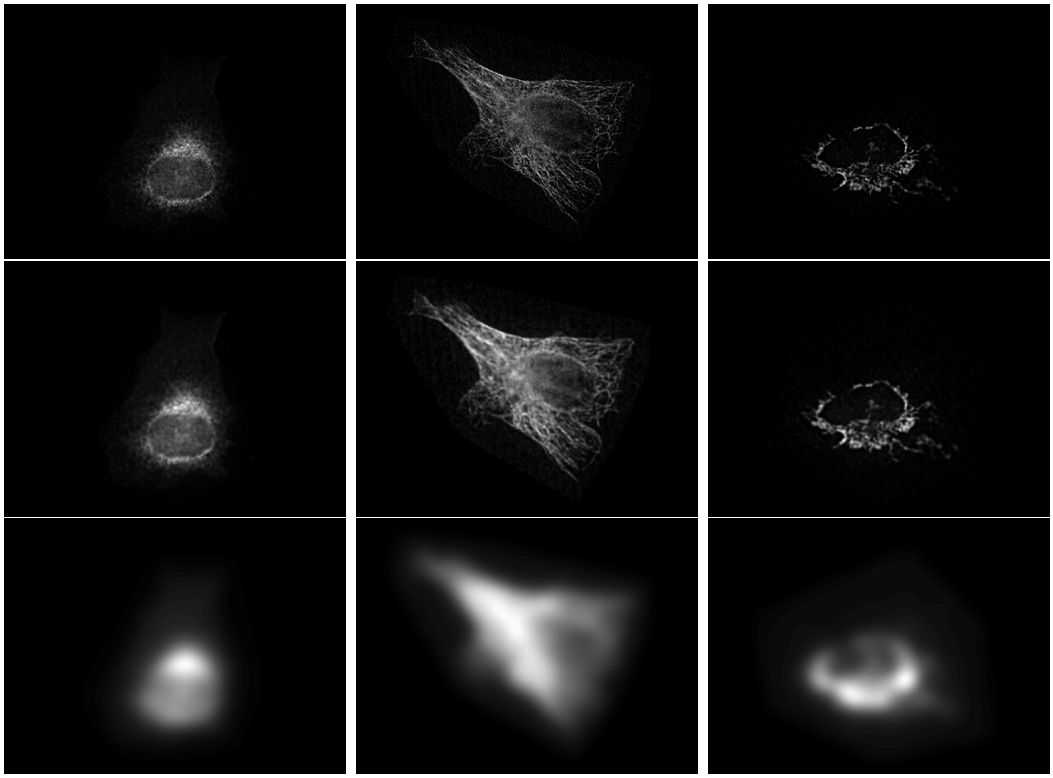


Figure 3.12: One cell example from each of the three classes. From left to right: ER, Microtubules, Mitochondria. From top to bottom: original images; blurred images with $e^{-\frac{1}{\sigma^2}} \approx 0.5$; blurred images with $e^{-\frac{1}{\sigma^2}} \approx 0.99$.

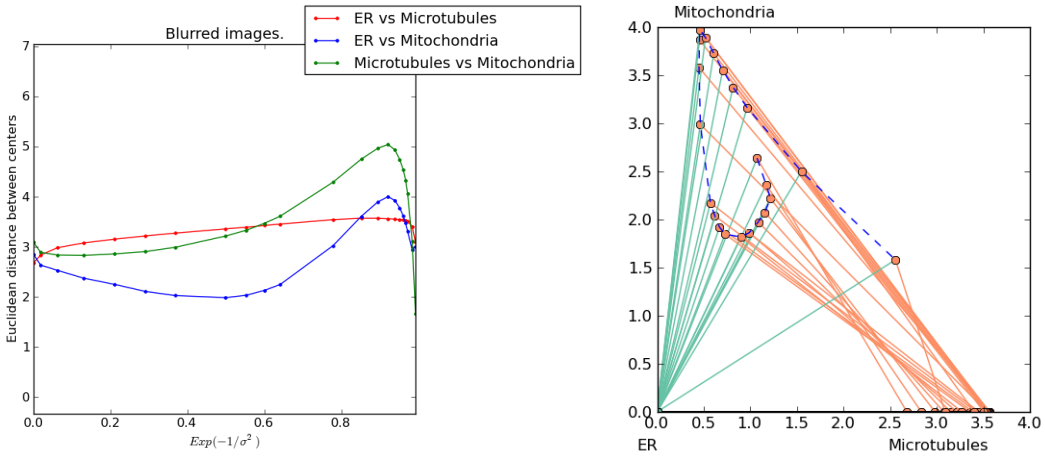


Figure 3.13: Blurring introduced to the protein data set. Left: scatter plot of pairwise centroid distances of three classes. Right: according triangle plot of pairwise centroid distances.

x-axis is the values of $e^{-\frac{1}{\sigma^2}}$, and the most left points from the original observed images (no additional blurring were applied). And we could observe that for the original data sets, the pairwise centroid distances of the features for the three classes are very similar with each other, but when the images were more and more blurred, the distances went to different ways: in a certain range, centroid distance between ER and Microtubules increases, and centroid distance between ER and Mitochondria decreases, while centroid distance between Microtubules and Mitochondria keeps approximately constant.

Right panel of Figure 3.13 is a triangle plot, each edge of a triangle represents the centroid distance of a pair of classes. We can see that for the original features, centroid distance between ER and Microtubules and centroid distance between ER and Mitochondria are very similar with each other, and are longer than the centroid distance between Microtubules and Mitochondria. When the images were more and more blurred, the centroid distance between ER and Mitochondria increases and achieved its peak at when $e^{-\frac{1}{\sigma^2}}$ is about 0.5, and the triangle appears like an approximate isosceles triangle at the beginning, then each centroid becomes further and

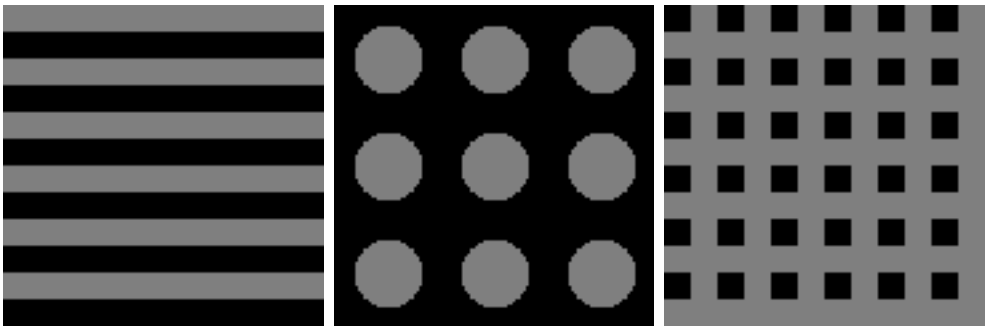


Figure 3.14: Binary images of three artificial patterns. From left to right: lines, circles, grids.

further apart, and after certain amount of blurring, the distances convergence.

As a supportive aspect of SIMEX-like approach, we apply this method to constructed artificial examples thereafter. We started with generating very clear and sharp binary images for three different patterns: lines, circles and grids and each pattern represents a class, which is shown on Figure 3.14.

Similar procedure are applied to the artificial images: using Gabor filters to extract conditional features, and then calculate the Euclidean distances between each pair of feature vectors. The results is shown in Figure 3.15, and we can see that within certain range of magnitudes of blurring, the linear trend exists, and when σ goes larger such as when $e^{-\frac{1}{\sigma^2}} \geq 0.6$, the distances still decrease following a linear model but with faster speed, i.e. large absolute slope. We also made a triangle plot as three of the pairwise centroid distances can just form a triangle, which is the right panel of Figure 3.15. When looking at the correspond triangle plot, the centroid distance between lines and grids and the centroid distance between lines and circles are very close to each other, no matter under which scale of σ value, thus we can always observe approximate isosceles triangles via σ changes, which indicates that the slopes of linear model for these two distances are very similar.

to infinity as $e^{-\frac{1}{\sigma^2}}$ goes to 1 ($e^{-\frac{1}{\sigma^2}}$ is an increasing function of σ), the distances converges to 0, which is consistent as the real data in Figure 3.13.

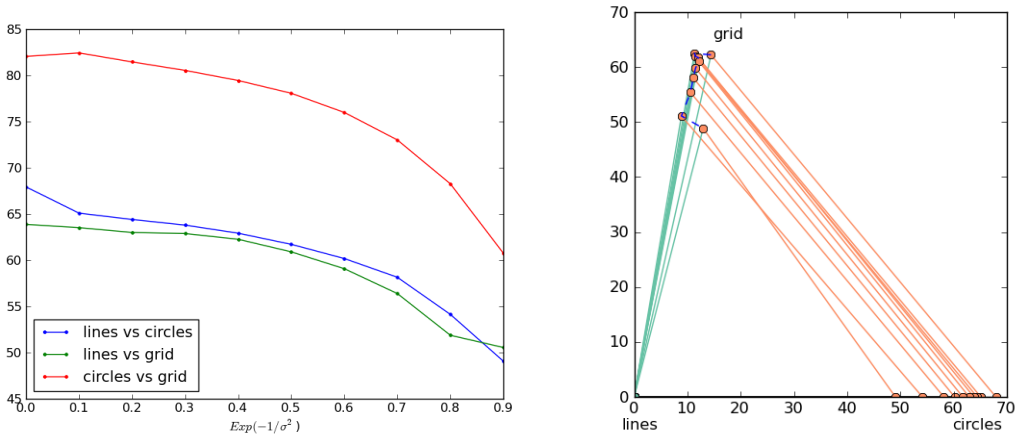


Figure 3.15: The pairwise centroid distances of artificial images when blurring were introduced. Left: scatter plot of pairwise centroid distances. Right: according triangle plot of distances.

Additive Noise

When introducing additive noise to images, i.i.d. normal errors with mean 0 and varied standard deviations are used. Again, we chose ActinFilaments, Endosome and Nucleus three classes from the protein data set. The result is shown in Figure 3.16. We can see that linearity appears in the left panel Figure 3.16, specifically, when the standard deviation of additive random additive noise is not very large, respectively speaking, for example, when the standard deviation of additive noise is small than 500, the pairwise centroid distances decrease linearly, eventually, the distances decrease linearly as well when standard deviation increase and go up to 3000, but with a smaller absolute slope. The overall slopes of linear model for three pairs of classes are not large number, which is also reflected by the triangle plot of Figure 3.16.

We again applied the procedure to artificial images, the patterns of artificial images are similar as shown in Figure 3.14, the only difference is that the none zero part in the binary images of Figure 3.14 is replaced by i.i.d. random normal variables (with mean 0 and standard deviation of 10) instead of value 1. The pairwise centroid distances are shown in Figure 3.17. We can conclude that when random normal additive noise

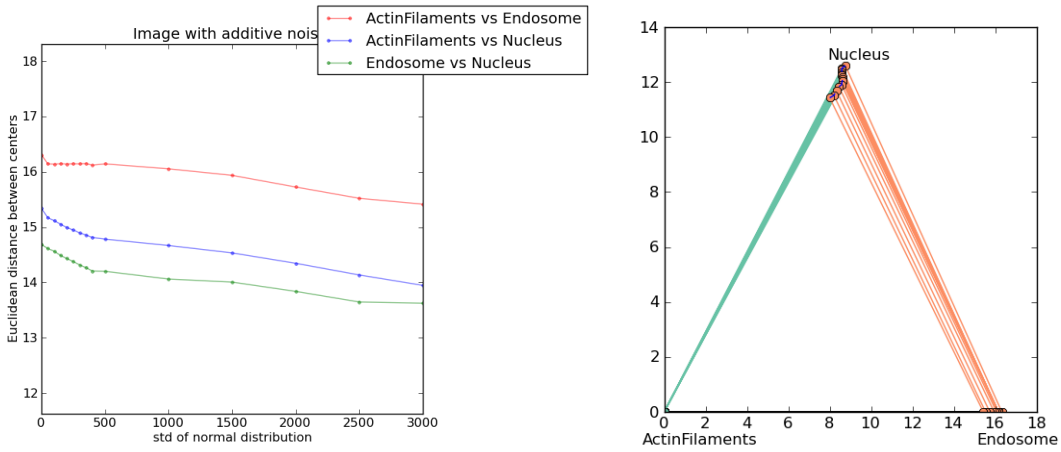


Figure 3.16: When additive noise introduced to the protein data set. Left: pairwise centroid distances of three classes. Right: triangle plot of distances.

included, the pairwise centroid distances change linearly in somewhat degree, and they can even be considered as a constant. The reason behind this fact might be that

Saturation

Finally, we considered the saturation issue in the protein data set. To model how the pairwise centroid distances change via different thresholding value which determines the degree of saturation, we used a list of thresholding values, and artificially increased the degree of saturation by choosing a threshold T and setting every pixel with intensity above T to be equal to T, therefore, the degree of saturation is a none-increasing function of threshold value T.

To best select this list of thresholding values for the protein data set, we made a box plot of maximum pixel number of each cell for each class, as depicted in Figure 3.19, we could see that the maximum pixel values for each class varies a lot, therefore, we chose to analysis ER, Mitochondria and Nucleus three classes alone which have similar maximum pixel values, as depicted in Figure 3.20. Figure 3.18 gives an example of original and saturated images for each of these three classes, respectively, from left to right: ER, Mitochondria and Nucleus; from to to right: original images,

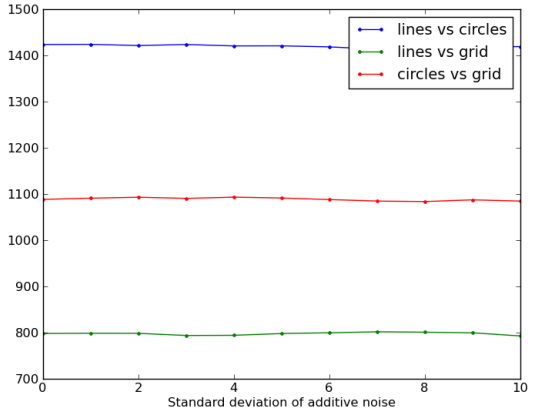


Figure 3.17: Additive noise: the distances between feature vectors of artificial images.

saturated images with $T = 80$, saturated images with $T = 60$, saturated images with $T = 40$.

The pairwise centroid distances are shown in the left panel of Figure 3.21, and the most right points give the centroid distances from raw images without artificial additional thresholding. We can observe that within certain range ($T \geq 200$), the pairwise centroid distances stay approximately constant, and when T continues to decrease, the distances slightly become smaller and keep increasing dramatically when the images were largely saturated. This conclusion can also be reflected from the triangle plot of Figure 3.21, for relatively large thresholding values, the corresponding triangles are almost identical, they are even hard to distinguish from each other in the plot, then the lengths of each side of triangles monotonely increase when smaller thresholding values were applied.

For the artificial images, we used the gray scale images with similar patterns as in Figure 3.14, the range of the pixels values of these gray scale images is from 0 to 1, and the images are are shown in Figure 3.22. The centroid distances are plotted in Figure 3.23. For the centroid distance between lines and grid, it is hard to summarize the pattern when the degree of saturation is not very large (such as when $T \geq 0.7$), but when T goes smaller (such as $T \leq 0.5$), the centers of these two images departed

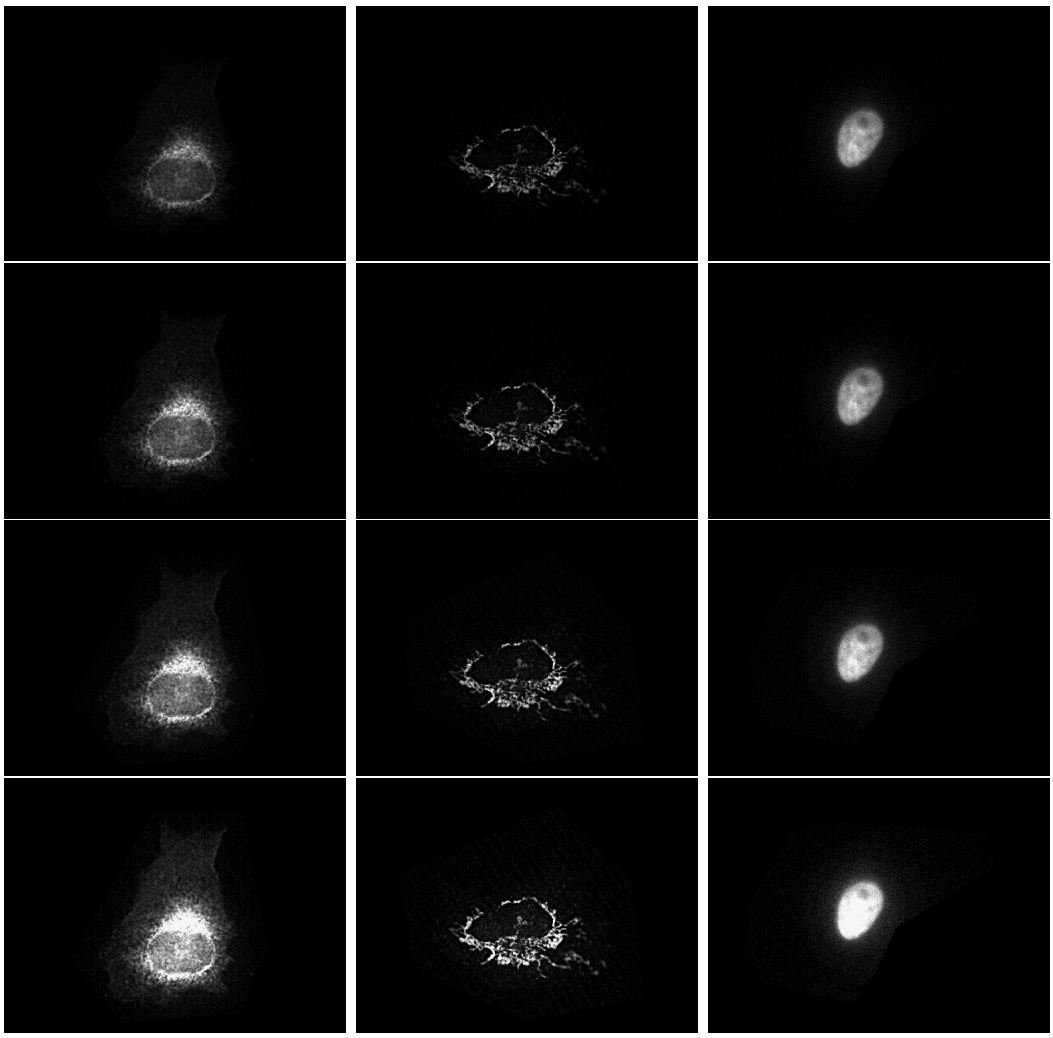


Figure 3.18: An example of original and saturated images for each of these three classes, respectively, from left to right: ER, Mitochondria and Nucleus; from to to right: original images, saturated images with $T = 80$, saturated images with $T = 60$, saturated images with $T = 40$.

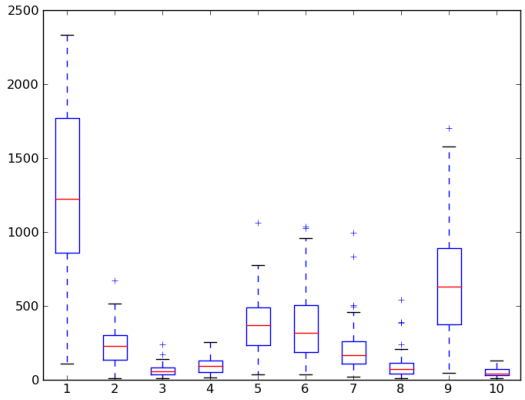


Figure 3.19: The box plot of maximum pixel value of each cell for each class in the protein data set. class 2: ER, class 8: Mitochondria, class 10: Nucleus.

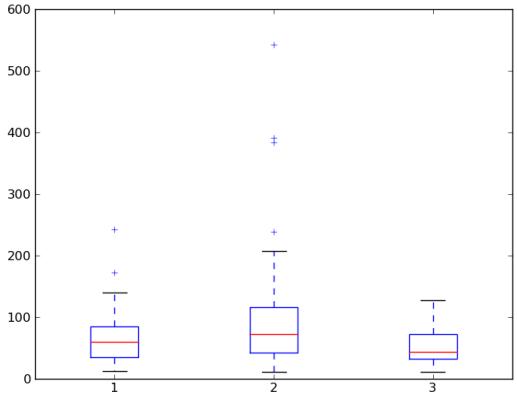


Figure 3.20: The box plot of maximum pixel value of each cell for three class in the protein data set. class 1: ER, class 2: Mitochondria, class 3: Nucleus.

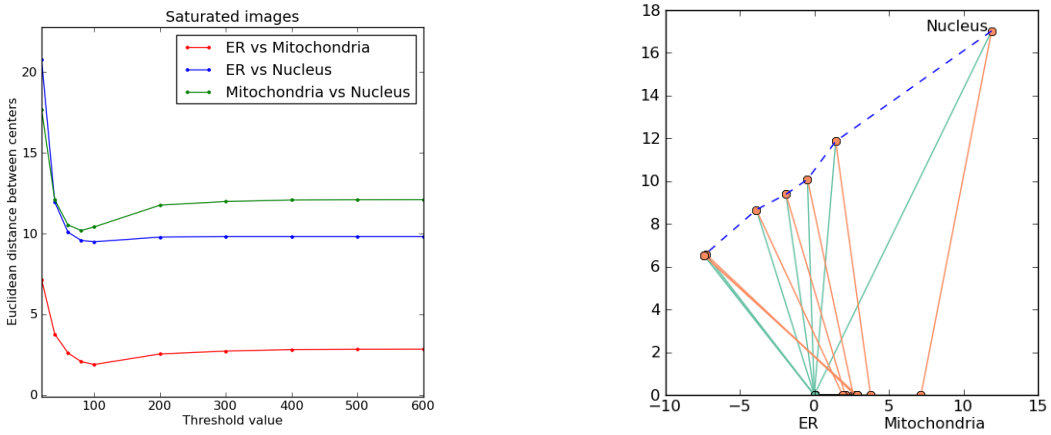


Figure 3.21: When saturation introduced to the protein data set. Left: pairwise centroid distances of three classes of the protein data set. Right: triangle plot of pairwise centroid distances.

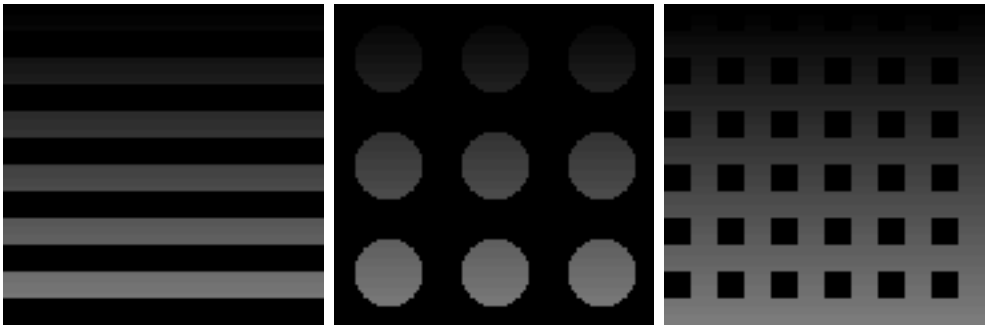


Figure 3.22: Gray scale images of three artificial patterns. From left to right: lines, circles, grids.

further and further linearly. And for the centroid distance between circles and grids, the overall trend appears approximately linear with significate nonzero slope. The other distance, which is between lines and circles, changes not dramatically.

3.3 Measurement errors for statistics capturing eccentricity of an elliptic distribution

We next considered a somewhat more complex statistic that aims to capture the extent to which a collection of points in high dimensional space is spherical versus

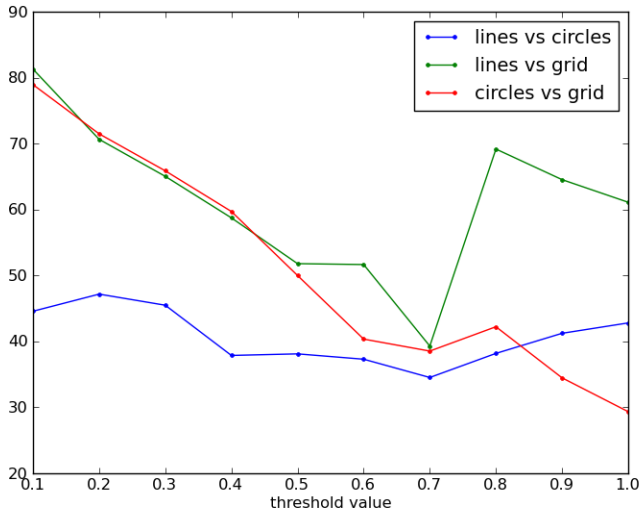


Figure 3.23: Pairwise centroid distances of three artificial patterns.

being eccentric (i.e. elongated on a particular axis). This property can be easily quantified by looking at the ratios between the dominant eigenvalue of the covariance matrix, and other eigenvalues such as the second eigenvalue, or the smallest eigenvalue. Again, we can ask whether a particular form of artifact in the images systematically affects the plug-in estimates of these measures.

Blurring

We use λ_1, λ_2 and λ_n to present the largest, second largest and smallest eigenvalue of the covariance of feature matrix, respectively.

We first address how $\frac{\lambda_1}{\lambda_2}$ and $\frac{\lambda_1}{\lambda_n}$ change via σ from the Gaussion fuction

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}.$$

The results are plotted in Figure 3.24, we could see that $\frac{\lambda_1}{\lambda_2}$ basically changed linearly (some classes even stayed almost constant) when the images became more and more blurred expect for the very extreme end (such as when $e^{\frac{-1}{\sigma^2}} > 0.9$). The value of $\frac{\lambda_1}{\lambda_n}$ have a linear trend within certain range only for four classes (ActinFilaments, Golgi

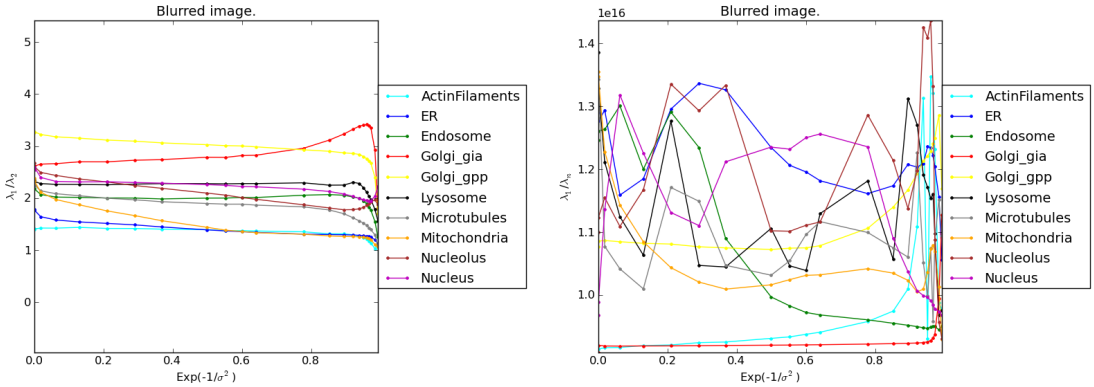


Figure 3.24: The degree of eccentricity of an elliptic distribution with blurring introduced to the protein images. left:$\frac{\lambda_1}{\lambda_2}$, right:$\frac{\lambda_1}{\lambda_n}$.

gpp, Golgi gia and Mitochondria), and for the other classes, it varies a lot and does not have a regular pattern. These plots demonstrated that the ellipticity represented by the ratio of the first two dominant eigenvalues for each class is a statistic which is barely depend on the degree of blurring of the protein images, therefore, if we extrapolate back to estimate the unobservable 'true images' , we would obtain similar ellipticity.

Moreover, for the protein data set, we applied PCA plot of feature-level data. The procedure is as described in Subsection 3.1.2 and Section 3.2. We selected the according PCA plot of ActinFilaments, Endosome and Nucleus these three classes as an example to visualize the change of eccentricity in two-dimension space, which is shown in Figure 3.25. From this figure, we are able to notice that when no blurring is present, these three classes are well separated, with very different degrees of eccentricity. When the images become more and more blurred, the distances between the centroids change, as do the degrees of eccentricity. When the images are extremely blurred, the pixel values are very close to each other, which indicates no obvious different localization distribution exist, therefore, at the end of Figure 3.25 (extremely blurred), the pairwise centroid distances are very small, and the degrees of eccentricity from three classes become much similar as well.

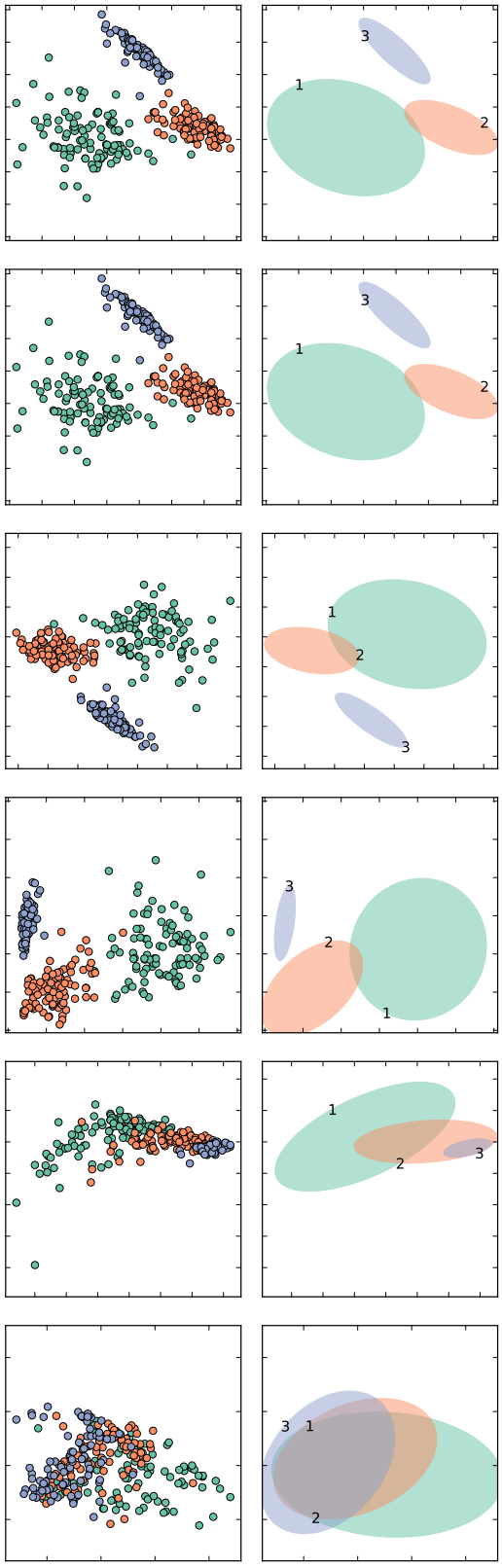


Figure 3.25: PCA of feature vectors for three classes with blurring. class 1: ActinFilaments, class 2: Endosome, class 3: Nucleus. Left: scatter plots. Right: Ellipses. From top to bottom: scale of blurring increases.

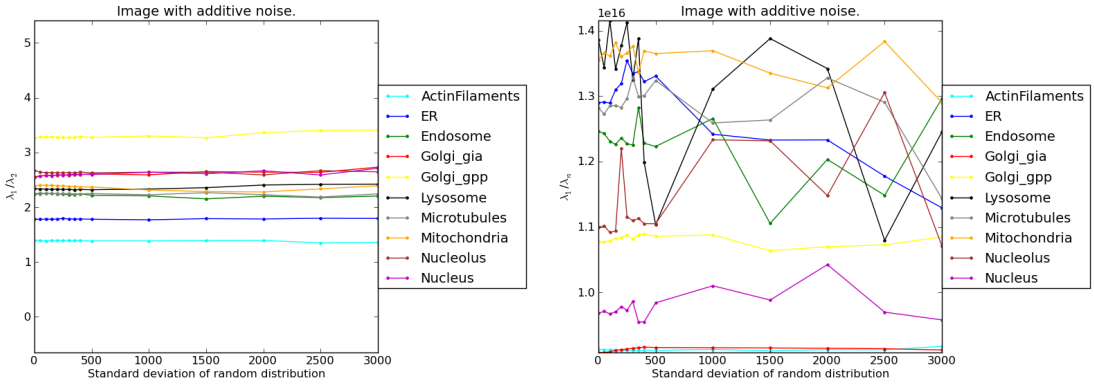


Figure 3.26: Additive noise: left:$\frac{\lambda_1}{\lambda_2}$, right:$\frac{\lambda_1}{\lambda_n}$

Additive Noise

Applying the similar procedure to the images with additional additive noise, we can obtain similar results as blurring: $\frac{\lambda_1}{\lambda_2}$ basically stayed almost constant, but $\frac{\lambda_1}{\lambda_n}$ does not show regular patten, except that three classes (ActinFilaments, Golgi gpp, and Golgi gia) have a linearly trend when the standard deviation of additive noise is not too large, as shown in Figure 3.26.

In addition, we are interested with the changes of eccentricity when additive noise introduced, therefore, we applied the PCA procedure to the protein data set, and again chose ActinFilaments, Endosome and Nucleus three classes as an example to illustrate the changes of the eccentricity in Figure 3.27. The set of standard deviations of additive noise applied in Figure 3.27 is a subset of standard deviations applied in Figure 3.16. From Figure 3.27 we can observe that when the standard deviation of additive noise is very large, with regard to the variation of feature level, the degree of eccentricity varies, but those changes are very small, not dramatic ones. In addition, the centroid distance between ActinFilaments and Endosome projected into two-dimension decreases with a small amount as well.

Saturation

As stated in Section 3.2, for saturation issue, due to different level of pixels values

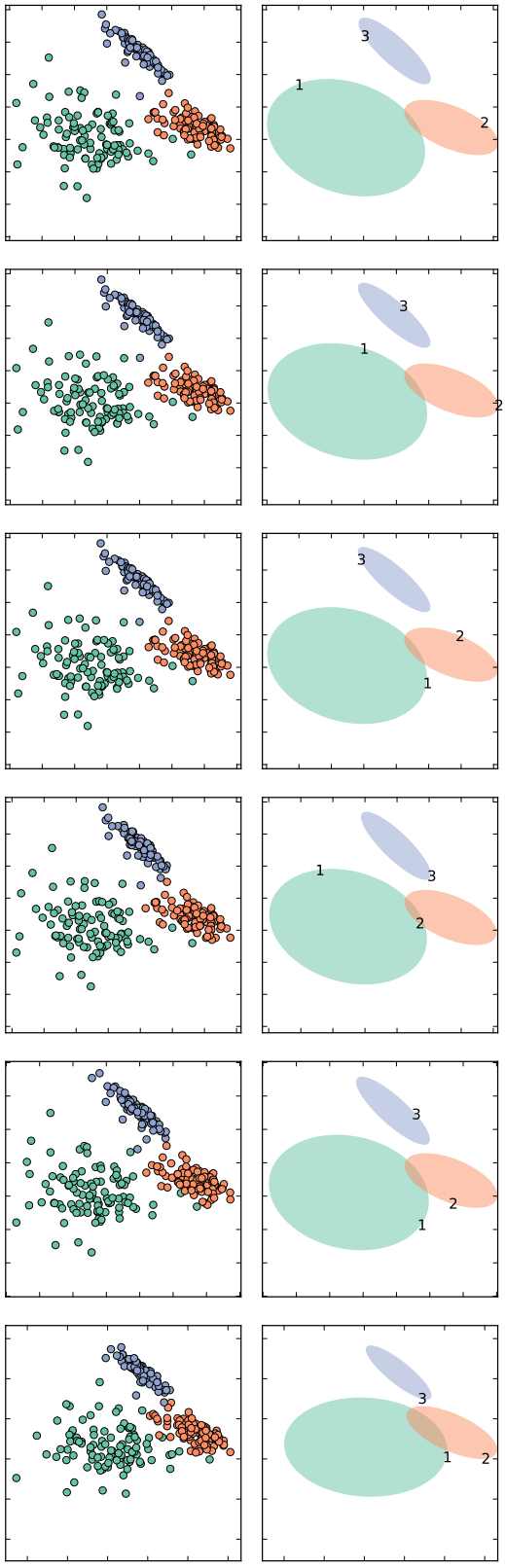


Figure 3.27: PCA of feature vectors for three classes with additive noise. class 1: ActinFilaments, class 2: Endosome, class 3: Nucleus. Left: scatter plots. Right: Ellipses. From top to bottom: standard deviation of random additive noise increases.

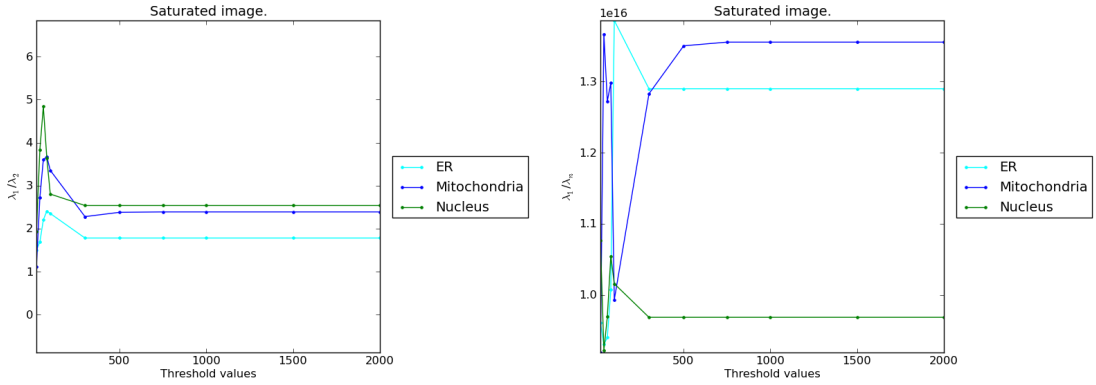


Figure 3.28: Saturation: left:$\frac{\lambda_1}{\lambda_2}$, right:$\frac{\lambda_1}{\lambda_n}$

in each class, here we only analyze ER, Mitochondria and Nucleus as well. Similarly, we drawn the plots of $\frac{\lambda_1}{\lambda_2}$ and $\frac{\lambda_1}{\lambda_n}$ as a function of T, which is shown in Figure 3.28. In this figure, the right end point is calculated from the original observed images without additional artificial saturation, therefore, images became more and more smoothed as T varies from left to right.

Based on Figure 3.28, especially the right panel of both plots when small magnitude of saturation were applied, we could have conclusion that both $\frac{\lambda_1}{\lambda_2}$ and $\frac{\lambda_1}{\lambda_n}$ basically stayed constant when the images were more and more saturated within a certain range. Therefore, the ellipticity of each class is a statistic which is barely depend on the degree of saturation of images as well, thus the ellipticity did not appear much bias on the observed images compared with the unobservable "true images".

From Figure 3.29, we notice that the degree of eccentricity projected into the first two dimensions change not very much when the strength of saturation is not strong, but it does appear show significant change, especially for class Mitochondria.

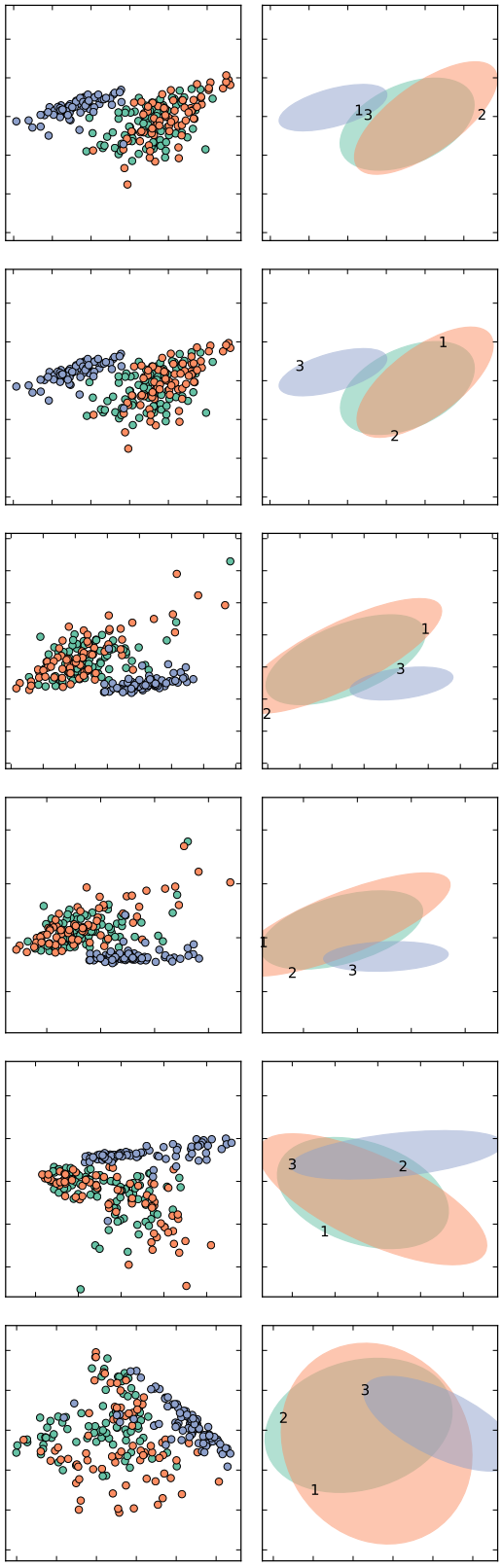


Figure 3.29: PCA of feature vectors for three classes with saturation. class 1: ER, class 2: Mitochondria, class 3: Nucleus. Left: scatter plots. Right: Ellipses. From top to bottom: thresholding values decreases.

3.4 Measurement errors for statistics capturing the angles between dominant axes of variation for different classes

We next considered another aspect of the distribution of images in PCA space, the angles between the dominate axes of variation for different classes, which aims to capture the rotation of features in high dimensional space. This statistic is quantified by looking at the relative angles between the dominant axes of the image feature covariance matrix for pairwise classes. Again, we can ask whether a particular form of artifact in the images systematically affects the plug-in estimates of these measures.

Blurring

For the protein data set, we again chose ActinFilaments, Endosome and Nucleus these three classes as an example, referring to Figure 3.25, we could notice that the relative angles between the dominant axes of the variation for pairwise classes change, and Figure 3.4 shows the specific numeric results, the degree of the angles in this figure is unit, and the range is from 0 to 180 as when we considering the relative angles of the dominant axes between classes, the direction of dominant axes is not considered. This also applies Figure 3.4 and Figure 3.4. When the images are moderately blurred ($0.1 < e^{-\frac{1}{\sigma^2}} < 0.6$), the relative angels change linearly, and such linear trend could possibly be increasing or decreasing.

Additive Noise

The example which illustrate the impact of relative angles of dominant axes of variation between classes by additive noise is also referring to ActinFilaments, Endosome and Nucleus, and Figure 3.27 shows how these classes move and expended (or elongated) in the PCA projection space. From Figure 3.27, we notice that the relative angles barely change, which can be reflected by Figure 3.4 as well. Figure 3.4 is a scatter plot with numeric results of relative angles of dominant axes between classes. We can observe that this statistic change linearly except that when the images were

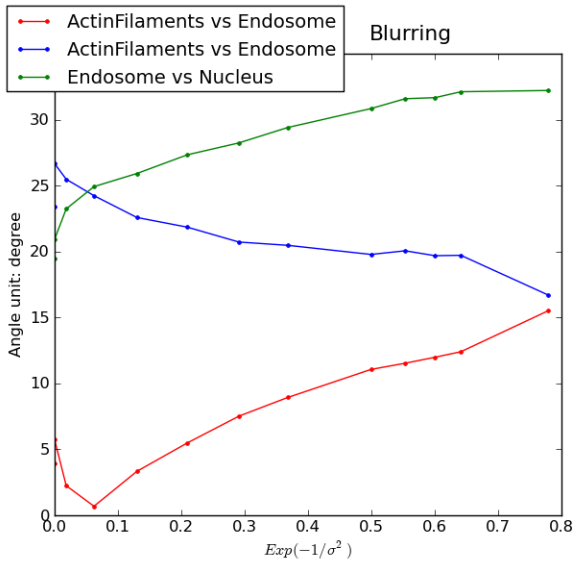


Figure 3.30: Pairwise relative angles for blurred images in three classes of the protein data set.

extremely saturated (when thresholding value is 20, which is shown at the very left end of the figure).

Saturation

As we stated in Section 3.2, the scales of maximum pixel value within each category vary a lot, therefore, we chose ER, Mitochondria and Nucleus as an example, similarly with Section 3.2 and Section 3.3. The PCA plot is Figure 3.29, the relative angles of dominant axes change a lot when the images are saturated, but it does not show regular pattern in this figure, therefore, we look at Figure 3.4, which shows numeric results. When the strength of saturation is low or moderate, it is difficult to estimate the trends of the changes of relative angles of dominate axes, but they change linearly when the strength of saturation begins to be strong.

This Chapter gives examples of measurement of different statistics under several sources of artifacts: blurring, noise and saturation, for both real data and artificial images, we then can have some conclusions that how the statistics (the pairwise centroid distance, the degree of eccentricity of dominant two dimensions and the

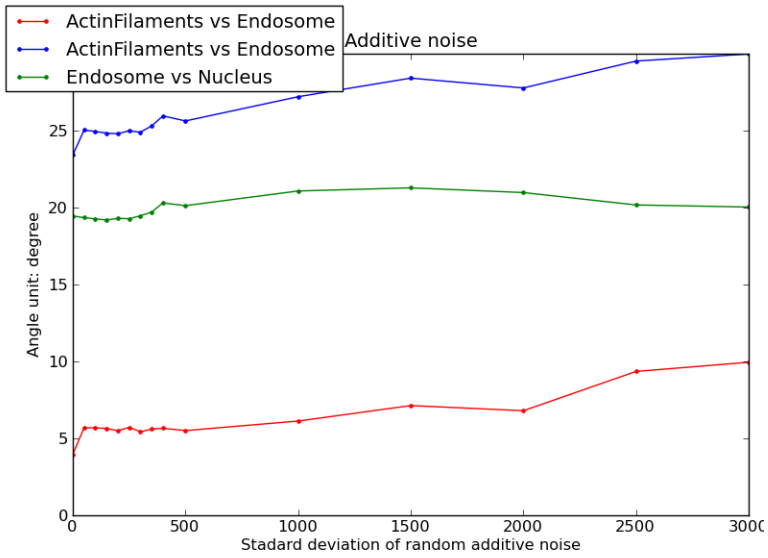


Figure 3.31: Pairwise relative angles for images with additive noise in three classes of the protein data set.

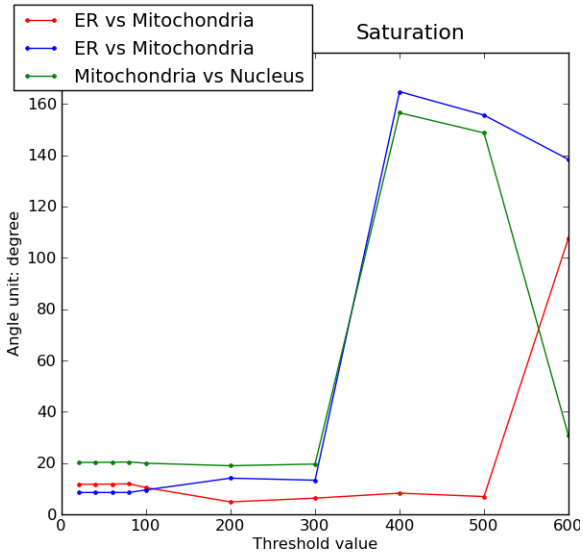


Figure 3.32: Pairwise relative angles for saturated images in three classes of the protein data set.

angles between dominant axes of variation for different classes) be influenced under different form of measurement errors. In some situations, we observe the changes of statistics is linear within certain range.

CHAPTER IV

Statistical analysis of timecourse image data from high content experiments

4.1 Introduction

Chapters 2 and 3 of this thesis considered several issues arising in the statistical analysis of image populations resulting from high content screening studies. These chapters were restricted to the analysis of single time-point images. High content experiments in which the wells are imaged repeatedly over time are also possible. Due to the exposure times involved in acquiring an image for each well on a plate, the repeated images are typically acquired at intervals ranging from several minutes to hours. Depending on the time scale of the phenomena being studied, this will often result in 5 to 20 images being acquired for each well.

Time course studies of cell cultures open up several new scientific directions. In the preceding chapters, we focused on experiments aiming to explore subcellular distribution patterns, either of proteins, or of small molecules. Here we will focus on small molecule studies, since small molecules exhibit much more variation in subcellular distribution over time compared to proteins. Specifically, we can consider the evolving patterns of subcellular staining from the moment that a compound is introduced into the cell culture medium, to the point that steady state distribution

is reached. This temporal distribution pattern is driven by the chemical kinetics governing the diffusion and active transport of a chemical inside a cell. For example, a compound that can rapidly pass through cellular membranes will reach its steady state distribution pattern more quickly than a compound that crosses membranes very slowly.

Time course experiments are subject to all of the artifacts discussed earlier for single timepoint studies. In addition, several other forms of artifacts specific to timecourse studies can also occur. Cells can grow, divide, or die over the duration that the images are being acquired. Cells can also move on the plate, so it is not straightforward to match cells in a well between timepoints. Limitations with resolution, contrast, and focus that were discussed earlier in the context of single timepoint studies are equally present in timecourse studies.

In this chapter, we focus on a specific, simple feature of timecourse images, namely, the degree to which the subcellular staining pattern is concentrated in or near the nucleus. This is a trait that evolves over time. A particular point of interest is whether different compounds accumulate in different regions at different times, as characterized in terms of their position in the cell relative to the nucleus. This characteristic can be quantified in images using a sequence of image processing steps. Our goal is to characterize the statistical performance of these steps, and to examine how this statistical feature of the image series is affected by image artifacts.

4.2 Quantifying subcellular staining patterns

As in earlier chapters, we will take advantage of the fact that the styryl small molecule image collection is acquired using Hoechst dye as a nuclear stain that is complementary to the staining resulting from the styryl molecules. Thus we have explicit measurements of the positions of the nuclear regions, although these measurements are subject to considerable noise and other artifacts. Our goal here is to

quantify the co-localization of the probe (i.e. the styryl molecule) with the nucleus, as reflected in the staining of the Hoechst dye. This can be accomplished by convolving the Hoechst image with an L_1 normalized disk-shaped filter with a given radius r. This has the effect of assigning to each pixel in the image a scalar value N that captures the net amount of Hoechst signal within an r pixel radius. We cap the Hoechst signal at a threshold T, so that pixels that are well within the nucleus have values $N \approx T$. Pixels roughly on the nuclear membrane have values N slightly less than $T/2$, depending on the curvature of the nuclear boundary. Positions more than r pixels from the nuclear boundary have small values of N, at the level of the background noise in the Hoechst images. Once N is calculated, we can now consider the conditional mean $E[S|N]$ of the signal S in the styryl channel. This conditional mean can be estimated using simple one-dimensional smoothing of the scatter plot of S against N.

The data set we analyzed are incubate live cells in 96 well plates with fluorescent small molecules. Co-incubate with Hoechst dye that labels the cell nuclei. Each image is a 512× 512 size image of cells at 18 time points following influx. The images capture signal from three different fluorescent wavelength bands and from the Hoechst band, at multiple exposure durations. Figure 4.1 shows one example of a series of high content screen images in time course experiment, so we could notice that some characteristic such the number of cells, the brightness and saturation issue varies across images. Figure 4.2 gives the correspond Hoechst channel images of Figure 4.1. Figure 4.3 represents the distances from nucleus for the pixels at the different positions at Hoechst channel, which could help us better understand the procedure stated above.

Plotting $\hat{E}[S|N]$, the estimated mean probe intensity, against N gives a curve which reflects the distribution of styryl channel intensity relative to the nuclei in the image. While in principal this procedure could be applied either to single cells

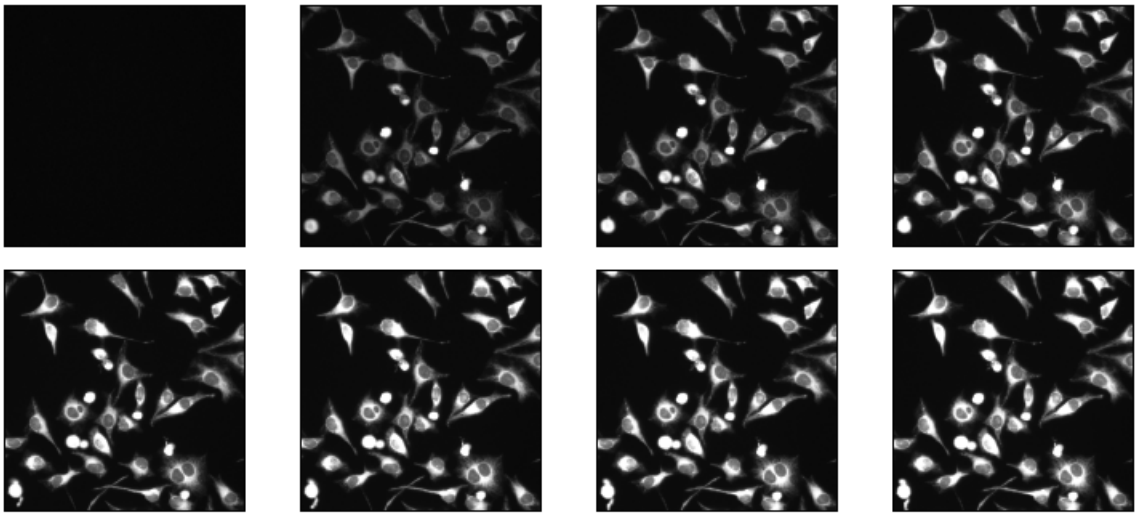


Figure 4.1: One example of a series of high content screen images in time course experiment.

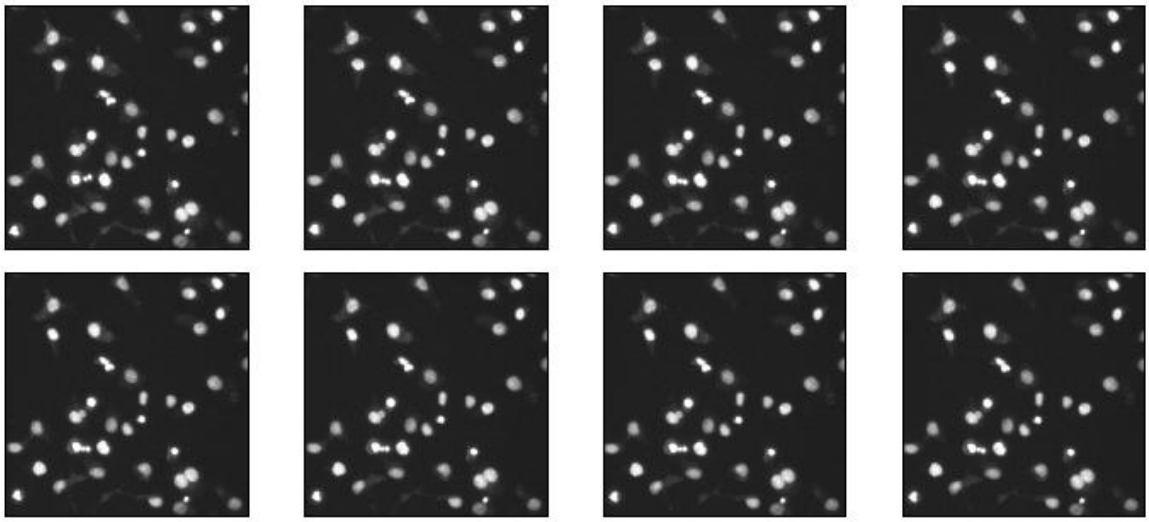


Figure 4.2: Hoechst channel images of Figure 4.1.

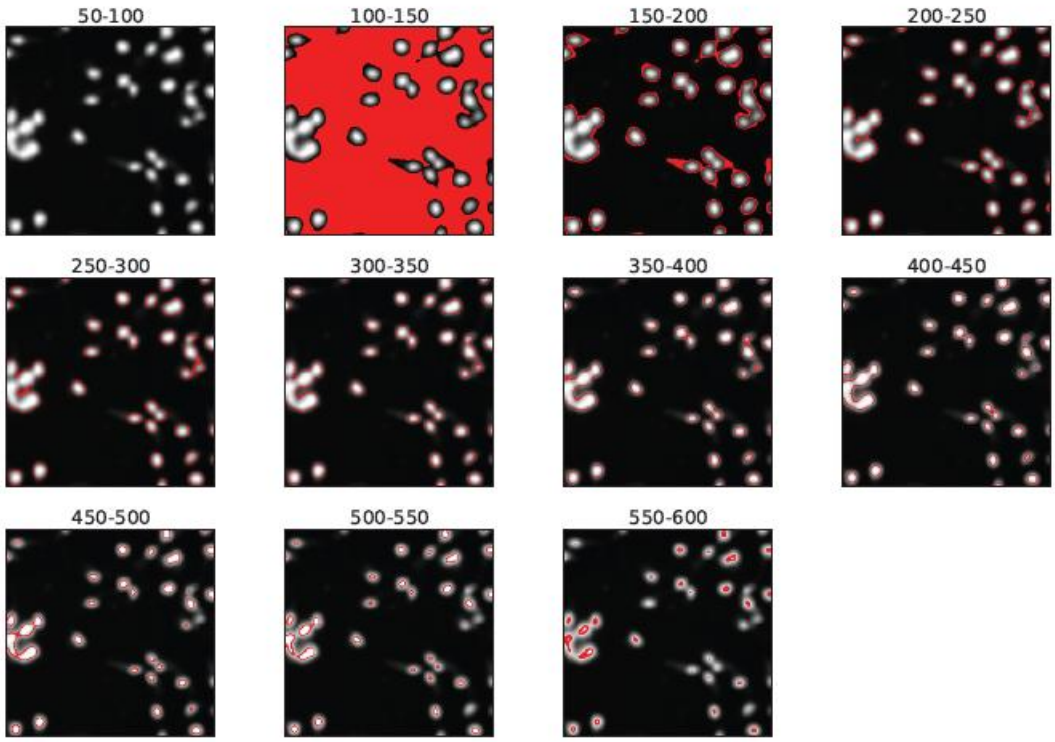


Figure 4.3: Distances from nucleus for the pixels at the different positions at Hoechst channel.

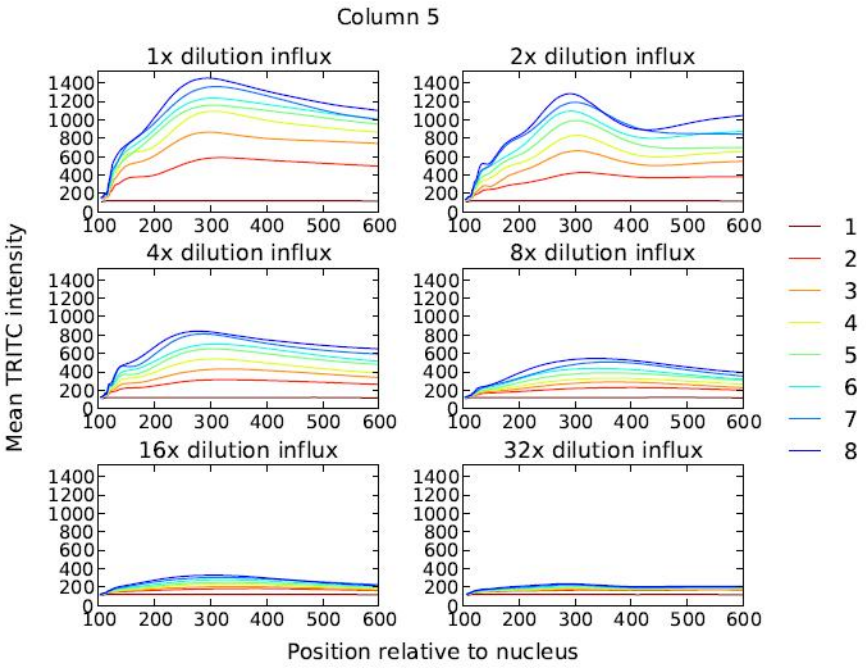


Figure 4.4: One example of estimated mean probe intensity by subcellular position.

or to images of entire wells, we only consider the latter case here. One example of estimated mean probe intensity by subcellular position are shown in Figure 4.4, the numeric labels in the right panel of these two figures give the label of exposure time. We then observe that for some time points and some positions, estimated mean probe intensity increases as the pixels go further from the nuclei, and when it achieves to the peak, it drops slightly down and stay almost constant, but when the pixels are far enough from the nuclei, or when the exposure time of cells is too short, the correspond estimated mean probe intensity then barely depends on the distance from nuclei, this can also be reflected from Figure 4.1, when the exposure time is too short, the cells are basically not been captured yet. In addition, we can observe that the positions of pixels where the estimated mean probe intensity arrives the peak are very similar from Figure 4.4.

When considering how $\hat{E}[S|N]$ changes via exposure time, one example is given in Figure 4.5, the numeric label at the right panel of this figure represents different

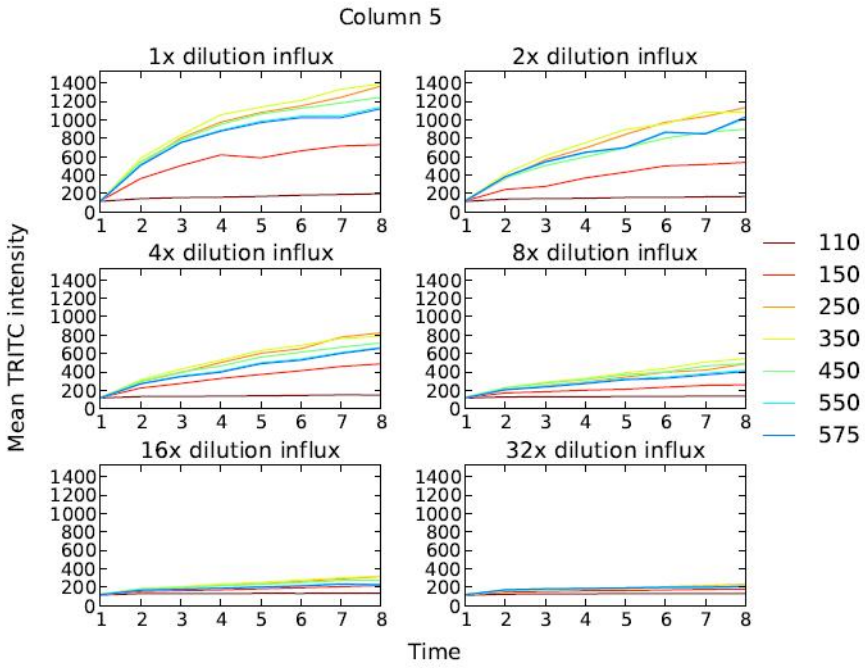


Figure 4.5: One example of estimated mean probe intensity by time.

positions at the Hoechst channel. If the exposure time is too short to capture any cell, of course $\hat{E}[S|N]$ would not change, so we only consider the situations that cells are captured. Basically, the estimated mean probe intensity would increase via exposure time increase.

4.3 Sensitivity analysis of extreme points affected by the introduction of image artifacts

Of particular interest in these plots will be the maximum and minimum points of $\hat{E}[S|N]$. The positions and heights of these points can then be considered relative to time, dose, or some other experimental factor. In a sensitivity analysis, we can also consider how these extreme points are affected by the introduction of image artifacts.

Blurring

Gaussian blurring is applied, similar as in Chapter 3. Figure 4.3 gives an examples

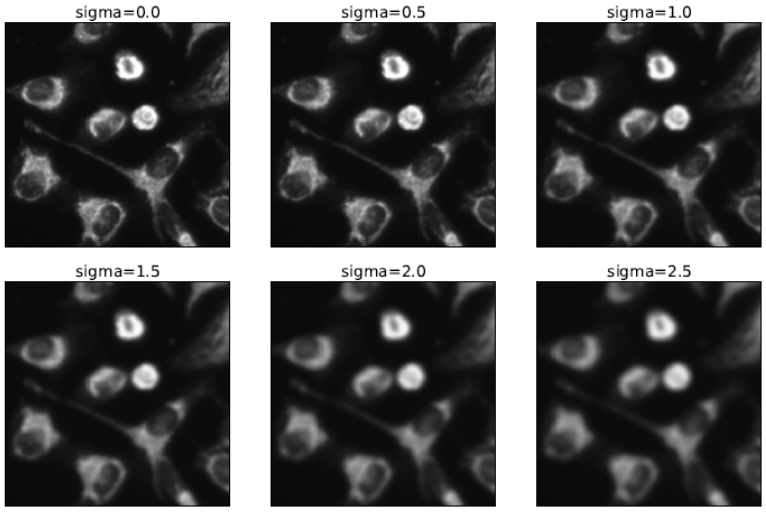


Figure 4.6: One example of images without manual blurring and with different strength of blurring.

of images without manual blurring and with different strength of blurring by choosing different σ values of Gaussian function.

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

Figure 4.3 gives an example of estimated probe intensity by subcellular position for original images(right panel, same as Figure 4.4) and for images that are blurred by $\sigma = 3$ (left panel). From this figure, we could notice that overall, the local maximum value and local minimum value of $\hat{E}[S|N]$ are somewhat influenced by blurring, especially for the first three dilution influx subplot. Be more specific, Figure 4.3 gives a more clear plot of estimated probe intensity by subcellar position for original image ($\sigma = 0$) and a series of blurred images ($\sigma = 1, 2, 3$). We notice that for the first local maximum point (when the label of position is around 150), the values of $\hat{E}[S|N]$ goes slightly up when the images are more and more blurred, but for the global maximum point (when the label of position is around 280), the values of $\hat{E}[S|N]$ changes oppositely: it becomes smaller and smaller when images are more and more blurred.

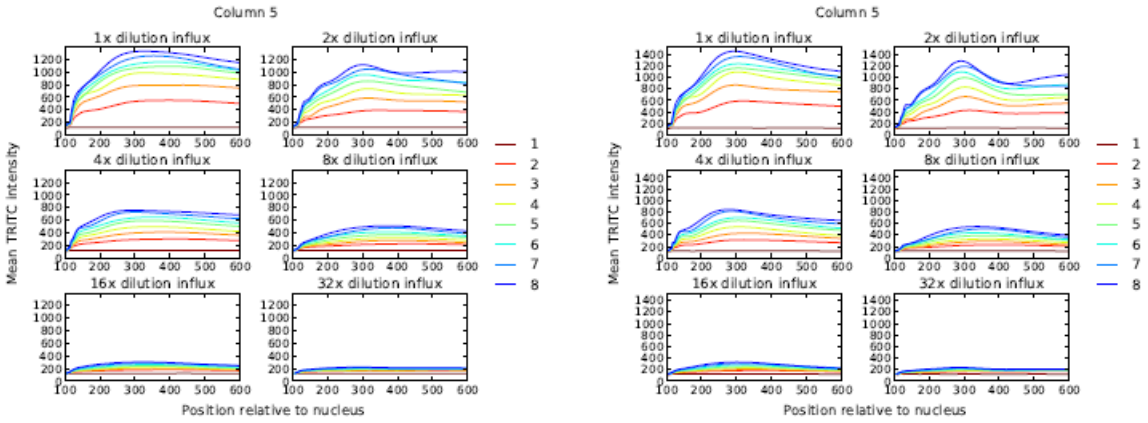


Figure 4.7: One example of estimated probe intensity by subcellular position for original image (right panel) and for images that are blurred by $\sigma = 3$ (left panel).

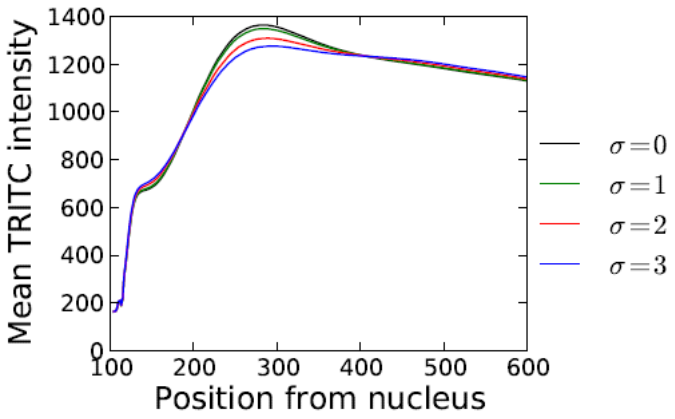


Figure 4.8: One example of estimated probe intensity by subcellular position for images introduced by different strength of blurring.

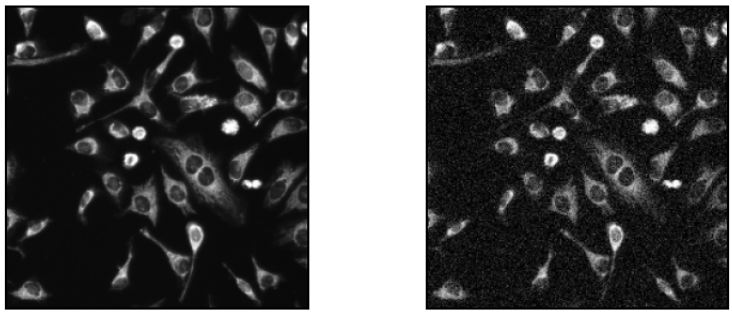


Figure 4.9: One example of observed image (left) and its correspond version with additive noise(right).

Additive noise

By adding random normal noise with mean 0 as similar as in Chapter 3, Figure 4.3 gives an example of images with additive noise, the left panel is the original observe image and the right panel is the correspond image with additive noise. Figure 4.3 gives an example of estimated mean probe intensity by subcellular position for original image (right panel) and for the images that are added with random noise. By comparing the two subplots in Figure 4.3, we hardly observe difference. This is expected since when random noise is added to the image, the mean probe intensity will not change due to mean 0 of additive noise, thus we see little difference for estimated mean intensity between original image and image with additive noise.

From this chapter, we considered whether different compounds accumulate in different regions at different times, as characterized in terms of their position in the cell relative to nuclei, and observed that the estimated mean probe intensity change by the distances from the nuclei and time, basically, it increases to its peak and then drops down slowly or approximately stay constant when the compounds are farther and farther from the the nucleus. It continue to increase when the cells are exposured longer and longer. We then explored how this statistic feature of the image series affected by image artifacts, and found out that additive noise does not influence

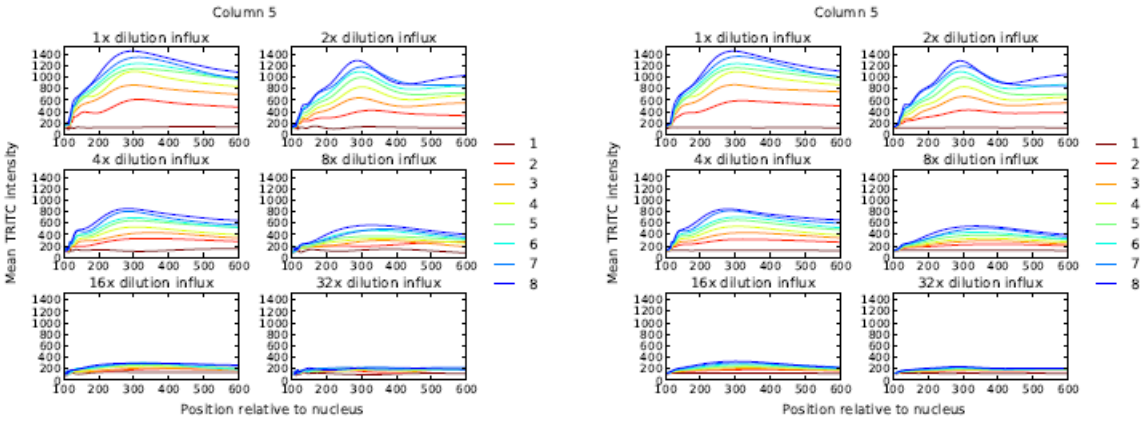


Figure 4.10: One example of estimated probe intensity by subcellular position for original image (right panel) and for images that is added with random noise.

much for the mean probe intensity as the mean value of additive noise is 0 and no nonlinear transformation involved when obtaining the image feature. Blurring has impacts on estimated mean intensity by subcellular patterns, especially the extreme points, and the trend can go either downward or upward.

CHAPTER V

Conclusions and Future Directions

In this thesis, we started with introducing the background information and motivation of analyzing high content screening images. Novel statistical approaches to handle HCS images are proposed in Chapter 2 through 4.

In Chapter 2, we mainly are interested with classifying subcellular localization of spatial distributions, therefore, we proposed a set of conditional features of quantiles of pixel contrast conditioning on pixel intensity after nonlinear filtering as image features. To support that the set of conditional features is a reasonable classifier and are more effective of marginal features, we started with constructing classes of artificial cells with different degrees of dependence between pixel values, then applying the procedure of extracting conditional features to those artificial cells, after statistical analysis, the result supports that the degree of spatial dependance is a major factor to distinguish two different distributions, in addition, the set of conditional features does perform significantly better than the set of marginal features. Thereafter, we applied similar procedure to three different HCS image data sets after certain low-level image processing, the styryl data set, the protein data set and the CHO data set, we again obtained that conditioning on pixel values is necessary as conditional features work better than marginal features.

For the styryl data set and the protein data set, we calculated the correspond clas-

sification rates with 95% confidence interval, and we did some regularization analysis for different types of parameters, including the classification method we used, the result supports that the baseline model we used, which is only based on the prior knowledge of cells without any optimization, is actually good at capturing the information from cells.

In Chapter 3, we are interested with the analysis of measurement errors, by extending the Simulation Extrapolation method from the setting of regression analysis with errors in variables to the setting of analyzing large image collections. The major sources of artifacts we considered are: out of focus of images, additive noise and saturation. The statistics we proposed are the pairwise inter-centroid distances of classes, the degree of eccentricity of variation projected into two-dimensional PCA space and the angles of dominant axes of variation for different classes projected into two-dimensional PCA space. The data we used including simple artificial random vectors without filtering, feature vectors of real images and feature vectors of artificial images. The basic procedure is that starting with the unperturbed data, we manually introduced different strength of the sources of artifacts to each data set, then made plots of how the statistics were affected by the artifacts.

For simple random vectors, we learned that linearity trend of each of the three statistics appears when each of the sources of artifacts is introduced, some characteristic, such as the pairwise centroid distances is even insensitive to the present of additive noise or moderated strength of truncation, this may due to its simple setting without any nonlinear transformation, such as Gabor filtering. For the features of real images and artificial images, the change of statistics become much more complex as nonlinear method of feature extraction were introduced, but we can still observe linear change within limited range of strength of artifacts, such as the change of pairwise centroid distances when saturation is not extremely strong. These observed facts support the use of linear approximations when applying the SIMEX procedure. In

addition, through these analysis, we learned that the degree of eccentricity presented by the length of first two dominant eigenvectors of the image population is relatively insensitive to the presence of blurring and additive noise, while pairwise centroid distances are biased either way of upward and downward when blurring is present.

In Chapter 4, we focused on the timecourse studies of high content screening images. When images are taken under different exposure time, then cells are possible to move, grow, divide, or die over the duration that the images are being acquired, therefore, we no longer analyze spatial distribution of localization of compounds in this chapter. A particular interest is whether different compounds accumulate in different regions at different times, this is carried on by proposing a simple feature of timecourse images to quantify the subcellular staining patterns, specifically, the degree to which the subcellular staining pattern is concentrated in or near the nucleus. We concluded that the estimated mean probe intensity change by the distances from the nuclei and time, basically, it increases to its peak and then drops down slowly or approximately stay constant when the compounds are farther and farther from the the nucleus. It continue to increase when the cells are exposured longer and longer. In addition, when exploring how this estimated mean probe intensity affected by image artifacts, we found out that additive noise does not influence much for the mean probe intensity and blurring has impacts on estimated mean intensity by subcellular patterns, especially the extreme points, and the trend can go either downward or upward.

In the future, we plan to develop methods for estimating the strength of various types of perturbations in real data sets, using calibration information; assessing other types of artifacts, such as foreign objects and debris, dead cells, and dye precipitates; assessing the effects of measurement errors on other types of analysis, such as cluster analysis and regression analysis of the images relative to chemical properties of the probes.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1]

[2] L. Beolchi and M. Kuhn. *Medical imaging: analysis of multimodality 2D/3D images*. IOS Press, 1995.

[3] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of Royal Statistics Society*, 1974.

[4] M. Boland and R. Murphy. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics*, 17:1213–1223, 2001.

[5] M. V. Boland, M. K. Markey, and R. F. Murphy. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry*, 1998.

[6] R. F. Boland, Michael V.and Murphy. A neural network classifier capable of recogniting the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics*, 17:1213–1223, 2001.

[7] H. Bunke and P. Wang, editors. *Handbook of character recognition and document image analysis*. World Scientific Pub Co Inc, 1997.

[8] California Institute of Technology. Making biological images sharper, deeper and faster. *ScienceDaily*, 2012.

[9] E. Candes and D. Donoho. Curvelets - a surprisingly effective nonadaptive representation for objects with edges. *Curves and Surfaces*, 1999.

[10] A. Carpenter and D. Sabatini. Systematic genome-wide screens of gene function. *Nature Reviews Genetics*, 5, 2004.

[11] R. J. Carroll and L. A. Stefanski. Approximate quasilikelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 1990.

[12] A. Chebira, Y. Barbotin, C. Jackson, T. Merryman, G. Srinivasa, R. F. Murphy, and J. Kovacevie. A multiresolution approach to automated classification of protein subcellular location images. *BMC Bioinformatics*, 2007.

[13] R. Chellappa and S. Chatterjee. Classification of textures using Gaussian Markov random fields. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1985.

[14] S. C. Chen, T. Zhao, G. J. Gordon, and R. F. Murphy. Automated image analysis of protein localization in budding yeast. *Bioinformatics*, 23, 2007.

[15] http://murphylab.web.cmu.edu/data/#2DCHO.

[16] C. M. Crainiceanu, B. S. Caffo, S. Luo, V. M. Zipunnikov, and N. M. Punjabi. Population value decomposition, a framework for the analysis of image populations. *Journal of the American Statistical Association*, 2011.

[17] C. M. Crainiceanu, R. Carroll, D. Ruppert, and L. Stefanski. *Measurement error in nonlinear models*. Chapman and Hall/CRC, second edition, 2006.

[18] T. N. Davis. Protein localization in proteomics. *Current Opinion in Chemical Biology*, 8:49–53, 2004.

[19] http://www.face-rec.org/general-info/.

[20] C. Faure and N. Vincent. Document image analysis for active reading. *'07 Proceedings of the 2007 International Workshop on SADPI*, 2007.

[21] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1984.

[22] S. Geman and C. Graffigne. Markov random field image models and their applications to computer vision. *Proceedings of the International Congress of Mathematicians*, 1987.

[23] L. Glesjer. Improvements of the naive approach to estimation in nonlinear errors-in-variables regression model. *Contemporary Mathematics*, 1990.

[24] R. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 1973.

[25] J. W. Hardin, H. Schmiediche, and R. J. Carroll. The simulation extrapolation for fitting generalized linear models with additive measurement error. *The Stata Journal*, 3, 2003.

[26] http://murphylab.web.cmu.edu/data/#2DHeLa.

[27] D. Hoiem, R. Sukthankar, H. Schneiderman, and L. Huston. Object-based image retrieval using the statistical structure of images. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 490–497, 2004.

[28] K. Huang and R. F. Murphy. Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics*, 2004.

[29] A. K. Jain and F. Farrokhnia. Unsupervised texture segmentation using Gabor filters. *Systems, Man and Cybernetics, 1990. Conference Proceedings., IEEE International Conference on*, 1990.

[30] R. E. Karlsen and G. Witus. Adaptive learning applied to terrain recognition. *Proc. of SPIE*, 6962, 2008.

[31] E. C. Kintner. On the mathematical properties of the Zernike polynomials. *Opt. Acta*, 1976.

[32] T. Kubota, T. Misu, T. Hashimoto, and K. Ninomiya. Image-based topographic recognition on natural terrain. *Proceedings of the 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2001.

[33] T. S. Lee. Image representation using 2D Gabor wavelets. *IEEE Transactions on pattern analysis and machine intelligence*, 1999.

[34] Z. Liu and C. Liu. A hybrid color and frequency features method for face recognition. *IEEE Transactions on Image Processing*, 17:1975–1980, 2008.

[35] S. Mallat. A theory of multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1989.

[36] B. Manjunathi and W. Ma. Texture features for browsing and retrieval of image data.

[37] M. K. Markey, M. V. Boland, and R. F. Murphy. Toward objective selection of representative microscope images. *Biophysical Journal*, 1999.

[38] N. Mittal, D. Mital, and K. L. Chan. Features for texture segmentation using gabor filters. *Image Processing and its Applications*, 1999.

[39] R. Murphy, M. Velliste, and G. Porreca. Robust classification of subcellular location patterns in fluorescence microscope images. *Proc 2002 IEEE Intl Workshop Neural Networks Signal Processing (NNSP 12)*, page 67, 2002.

[40] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 1996.

[41] U. Park and A. K.Jain. 3D model-based face recognition in video. *The 2nd International Conference on Biometrics*, pages 1085–1094, 2007.

[42] http://www.cellprofiler.org/.

[43] L. Stefanski and J. Cook. Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, 1995.

[44] http://www.geo.hunter.cuny.edu/terrain/.

[45] S. Venkataraman, J. L. Morrell-Falvey, M. J. Doktycz, and H. Qi. Automated image analysis of fluorescence microscopic images to identify protein-protein interactions. *Proceedings of the 2005 IEEE, Engineering in Medicine and Biology 27th Annual Conference*, 2005.

[46] M. Vetterli and C. Herley. Wavelets and filter banks: theory and design. *IEEE Transactions on Signal Processing*, 1992.

[47] Y. N. Wu, C. Guo, and S. C. Zhu. From information scaling of natural images to regimes of statistical models. *Quarterly of Applied Mathematics*, 2008.

[48] Y. N. Wu, S. Zhu, and D. Mumford. Filter, random field, and maximum entropy (FRAME): towards a unified theory for texture modeling. *International Journal of Computer Vision*, 1997.

[49] http://yeastgfp.yeastgenome.org/.

[50] V. Zipunnikov, B. S. Caffo, D. M. Yousem, C. Davatzikos, B. S. Schwartz, and C. M. Crainiceanu. Functional principal component model for high dimensional brain imaging. *NeuroImages*, 2011.