Title

Data-driven Computing and Analysis with Contrasting Statistical Developments in Real-world Applications

Author

Liao, Shuting

Publication Date

2023

Data-driven Computing and Analysis with Contrasting Statistical Developments in Real-world Applications

By

SHUTING LIAO
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOSTATISTICS

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Fushing Hsieh, Co-Chair

Debashis Paul, Co-Chair

Patrice Koehl

Committee in Charge

2023

To my family

Contents

Abstract

The real data generated from the real-world complex systems in general embraces rather sophisticated deterministic and stochastic structures on multiscale levels. Such structural complexity surely induces very challenging learning problems and poses very difficult data-analyzing issues. Data coming from diverse complex systems studied in scientific fields are often found to have diverse ways of preserving data pattern information. This diversity of ways of encoding information is in part due to the constraints between data's sophisticated deterministic and stochastic structures. It becomes necessary for data scientists to adapt to such sophisticated constraints by adopting data-driven computing approaches when analyzing data from real-world complex systems. That is, to gain authentic information in data, it is essential to develop data-analysis methodologies according to the data's intrinsic characteristics. In this dissertation, we develop and propose data-driven adaptive computational methods and statistical frameworks based on specific data structures, including digital images, data on Alzheimer's Disease as well as limited data on biochemical experiments. In a project of evaluating the effectiveness of chemical spraying through an unmanned aerial vehicle (UAV), we prescribe a computational approach to using color-identification algorithms and minimum spanning trees (MSTs) to analyze the spatial distribution of color dots of various sizes and colors on the image. We succeeded in achieving the goal of testing the evenness of mechanical spray via color-dot testing papers. In a project studying the aging effects on a series of three of Van Gogh's Sunflowers in a vase, we develop a computational approach to restore the original color and vibrancy in a reverse-engineering fashion. Their already faded or brownish-yellow backgrounds are successfully revived to shed yellow-oriented lights computationally. In a project of analyzing time-to-event data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, we employ conditional entropy to unravel heterogeneity among subjects and evaluate the potential factors that affect the diagnosis of Alzheimer's disease. Our data-driven results are compared to Cox's proportional hazard modeling and demonstrate better capability in identifying significant factors. In a contrasting fashion, we also study a statistical problem in modeling biochemical experiments with data being limited in size and scope. Under such constraints, we propose a flexible methodology for analyzing the variability of smooth functionals of the growth or production trajectories associated with temporally measured biochemical processes across different experimental conditions when the amount of data is limited. We demonstrate, through numerical experiments and real data analysis, the effectiveness of the statistical inference of key parameters of interest and the flexibility to extend to correlated structures. We conclude that data-driven approaches are necessary when analyzing big data sets, while statistical modeling has its merit when data is limited.

Acknowledgments

I would first like to thank my advisors, Fushing Hsieh and Debashis Paul, who have patiently guided me through my doctoral study. Throughout my years as a doctoral student, Profs. Hsieh and Paul have generously provided invaluable assistance to my efforts in training me as an independent and qualified researcher. With their guidance, I've been successfully studying and solving varying but all essential realistic and interesting problems. During the weekly meetings, they not only provide professional suggestions and guidance on my work but also always inspire and encourage me. Their mentoring has been instrumentally helpful. In addition, they kindly offered help in the revision of both manuscripts and presentations. I truly appreciate it to have the best two advisors in the world in my crucial years.

I also would like to thank Professors Fushing Hsieh, Debashis Paul, Thomas Lee, and Sanjay Chaudhuri for providing me with chances to work on fruitful interesting projects, which greatly enhanced my skills and understanding of my research. I derive a great deal of knowledge and skills from their invaluable and insightful instruction and assistance.

I'm thankful to Professors Fushing Hsieh, Debashis Paul and Patrice Koehl for serving as my dissertation committee members, reading the manuscript, and providing insightful advice. Also thanks to Professors Patrice Koehl, Alexander Aue, Christiana Drake, Somen Nandi, Miriam Nuno for serving on my qualifying exam committee and offering valuable comments.

Last but certainly not least, I would like to express my deepest gratitude to my parents and my boyfriend, Tuo Wang. I would never have these achievements without their support and love. Their love makes me strong and brave to pursue my dream.

CHAPTER 1

Overview

The amount and complexity of data have increased exponentially, with the emergence of complex questions of interest and new technologies. Due to the practical needs and new technologies that emerged, real data have more complex structures. The real data can be characterized by several possible interacting mechanisms that preserve collective behaviors in terms of temporal, functional and spatial structures. And the corresponding questions of interest become more sophisticated so that traditional methods fail to answer them. For instance, traditional models such as linear regression or random forest are not suitable to process and analyze images for the region of interest identification. Even when dealing with more traditional data, we can see more information and problems which can not be captured and solved by traditional methods. For instance, the results of the Cox model on time-to-event data are suspicious to be informative and reliable enough to discover the significant factors and uncover the heterogeneity. Biological experimental data, as another example, can not be analyzed through a standard linear regression framework due to the nature of non-decreasing growth trajectories. And the inference can be challenging because the data is usually limited because of time and cost restrictions. Therefore, it is imperative to be more cautious and careful to integrate information by developing a data-driven method that is capable to answer the data-specific questions of interest. We focus on several interesting real data sets with different questions and goals, by developing computational methods. The main goal is, to develop computational tools for researchers to uncover the information so that the data is better understood.

This dissertation focuses on various domains that require data-driven techniques to address real-world questions that traditional methods cannot answer. We analyze multiple datasets, each with its unique structures, limitations and scientific goals. Using scientific objectives and practical constraints as our guide, we derive computational methods aimed at the particular characteristics and goals of each dataset. In this dissertation, we demonstrate the essential role of data-driven approaches in leveraging information from complex systems and tackling practical questions with reliability.

1.1. Data Science Application in Color Science

One challenging problem is data science applications in color identification in images. Color digital images are usually taken by photos or scanned on paintings, and thus the color objects or regions can be characterized by information in terms of spatial positions and color. It is useful to extract and learn the target region of interest from a given image because the corresponding color and spatial information can be investigated and utilized. One application, as an example, is to test whether the purple dots on a rectangular yellow paper is distributed uniformly based on the image. Spray technologies via unmanned aerial vehicle (UAV) for liquid chemicals of fertilizers, herbicides, and pesticides are vital in Precision Agriculture due to economic and environmental perspectives. These technologies can save costs from many aspects and add capabilities of dynamic and optimal management. However, the success of such technologies heavily relies on effective evaluations of their performances in terms of efficiency and precision, such as testing the sprayed liquid droplets' homogeneity and spatial distribution on a target area through color image analysis. There are two main challenges: First, the sample size of pixels in the image can be too large for the computation. Second, it is crucial to characterize the spatial information of identified purple dots and conduct the testing, where theoretical statistical hypothesis testings are not working. Therefore, a data-driven approach is essential for this practical problem. Developing such a method for this problem will be helpful as it can aid to uncover more informative evaluations about the quality of spray technologies. This can serve as an insightful assistant tool for manufacturers to inspect and adjust their technologies at a low cost.

Such color study based on data science can also be beneficial in digital paintings, such as van Gogh's *Sunflowers in a vase with a yellow background series.* This is motivated by the fact that a large collection of digital scanning of these paintings is available to data scientists. More importantly, it is noted that researchers like museum curators mostly focus on chemical investigation and microscopic analysis, either through the pigments or the canvas, for the purpose of comparison among paintings as well as color restoration. Such analysis, however, is relatively time-consuming.

The color restoration of van Gogh's *Bedroom* by the van Gogh Museum, is realized after a six-month study. We aim to study the paintings from the data science perspective in color distribution, to learn the differences between the original version and its repetitions, as well as the aging pattern, through a computational framework. The core challenges are the large sample size of pixels contained in one image and compare the target regions. The first problem can be solved by the characteristics of color consistency. The second one can be solved by dividing the paintings into three main regions of interest and comparing their color distribution correspondingly across paintings. Such a framework will be promising because it is able to uncover the color information intended by van Gogh and reconstruct the aging color through effective computational algorithms. It is beneficial to serve as an assistant to help researchers have a better understanding of protecting paintings and preventing the aging effect.

In Chapter 2, we propose a computational method to realize the testing of purple dots in an image spatial-wise. One key step of this testing involves color image analysis consisting of two coupled computational tasks: exhaustive color-dots identification and spatial pattern extracting and testing. These two tasks are in accord with two natural perspectives of color complexity of a color image [**109**]. These two key problems can be resolved by making use of color consistency and the minimum spanning tree of local spatial squares We analyze five images of yellow water-sensitive papers that were exposed to aqueous spray droplets during drone experiments. Each paper was placed on a different background and the resulting images were captured under varying lighting conditions. We propose a computational method to extract precise spatial and distributional information of all blue-purplish dots and to prescribe their interacting relations as joint characteristics, characterize spatial geometries with respect to varying dot sizes as the targeted joint characteristics, and conduct spraying tests.

The analysis of the color image inspires us to develop a systematic computational approach to solving practical problems of the well-known paintings, *Sunflowers* by van Gogh, which is detailed in Chapter 3. Van Gogh's *Sunflowers* appeal to many. Scientists and museum curators have used a broad array of traditional and state-of-the-art techniques to look at and below the paint surface itself [**45**, **52**, **80**, **81**, **82**, **83**, **84**, **100**, **104**]. The differences between the original version and its repetitions have been studied by the van Gogh Museum [**46**]. In addition, they have been continuously studying and working on the aging pattern of paintings due to time and chemical factors. Those problem has been investigated in many ways. As the fact that digital images are accessible, we would like to contribute via a data-science-based approach to those problems. We utilize unsupervised learning techniques for color analysis. To further analyze the color information of the paintings, we also consider the color distribution within the sunflowers. To do this, we divide the sunflowers into smaller regions and compute the color distribution within each region. We then compare the resulting color distributions for the sunflowers in each painting, which gives us a more detailed understanding of the differences in color between the paintings. Overall, our approach to analyzing the paintings is complementary to the traditional methods used by scientists and museum curators. By using data science techniques to analyze high-resolution images of the paintings, we are able to gain insights into the color information of the paintings that may not be readily apparent from traditional analysis techniques.

1.2. Unraveling heterogeneity of ADNI's time-to-event data using conditional entropy

Apart from the digital image analysis, the data-driven approach paradigm can also be applied to the analytical study of Alzheimer's Disease Neuroimaging Initiative (ADNI) to learn about the heterogeneity in the time-to-event data and identify the significant factors. It is noted that Alzheimer's disease (AD) is the most common cause of dementia in older adults and it is irreversible neurodegenerative in a loss of mental function. However, the cause of Alzheimer's disease in most people has not been fully understood yet and thus identifying key factors leading to the development of the disease is crucial. The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease (AD). Numerous cross-sectional studies have utilized data from the ADNI database to track the progression of Alzheimer's disease (AD) with biomarkers. These studies commonly apply the Cox Proportional Hazard (PH) Regression model, a well-established statistical method for time-to-event data, such as the progression from CN to MCI or from MCI to AD. The partial likelihood approach, which is extensively studied, is often used as the primary inferential tool for identifying significant biomarkers. Nevertheless, this widely used approach in Survival analysis encounters several fundamental challenges when applied to real-world data. For instance, there is no guarantee that the Cox PH modeling assumptions on the non-informative censoring mechanism align with the pattern information embedded within the data. Moreover, we suspect that, given that heterogeneity among subjects is intrinsic in ADNI data, Cox PH is capable to provide reliable results, in particular when facing heavy censoring rates on the global and local scales and potentially complex interacting relations and effects among covariate features. Therefore, it is fundamental to propose a more appropriate data-driven method with fewer assumptions.

In Chapter 4, we focus on the major factor analysis of Alzheimer's Disease Neuroimaging Initiative (ADNI) data, by proposing a non-parametric framework. We aim to detect AD at the earliest possible stage (pre-dementia) and identify ways to track the disease's progression with biomarkers. We apply the Categorical Exploratory Data Analysis (CEDA) paradigm to evaluate conditional entropy-based associative patterns between the categorized response variable against 16 categorized covariates all with 4 categories. We also develop a display called conditional-entropy-expansion to unravel the effects of all chosen perspectives of heterogeneity.

1.3. Analysis of variability of functionals of recombinant protein production trajectories based on limited data and its extensions

Another challenging problem with the real data is, on the contrary, the limited size, which is commonly seen in clinical or biological experiments. Due to the limitation of time and resources, the size of units measured during the experiments can be extremely small. One example is conducting experiments to develop a cost-effective alternative to the human enzyme butyrylcholinesterase (BChE) found in blood plasma. It can be used as a prophylactic and/or therapeutic agent against nerve agent poisoning but is limited by its high cost. Rice-produced recombinant human BChE (rrBChE) has been developed as a potential alternative source of BChE, with transgenic rice cell suspensions being one of the host expression systems used to produce it [**3**, **74**]. The rice alpha-amylase 3D (RAmy3D) promoter regulates the production of rBChE in these transgenic rice cells, which are grown in a sugar-rich medium for production and transferred to a sugar-free medium for rrBChE production [**19**, **49**, **50**, **76**, **99**]. However, growing plant cell suspension cultures is both time-consuming and expensive, with the slow growth rate of plant cells being a major challenge. Transgenic rice cells take 10-12 days to be cultivated in batch culture [**19**, **76**]. When conducting experiments involving multiple factors or conditions, limited time for cultivation and equipment availability can restrict the number of bioreactor replicates. As a result, interpreting the data or selecting the best conditions may present challenges. Due to the high cost involved, it is important to gain a thorough understanding of the data. For instance, one key question of interest is how different it is in production trajectories across treatments. This can be answered statistically by making inferences on the related parameters from the modeling. Though there are admittedly kinds of traditional statistical methods available to quantify the production and answer inferential questions, they are not reliable in this situation because of the limited sample size. Hence any statistical inference procedure that directly relies on large sample theory will have limited accuracy or may be misleading. Moreover, since some of the parameters (functionals of the production trajectories) or process metrics of interest are nonlinear, the standard ANOVA framework that relies on the linear model theory does not apply. Therefore, it is imperative to aid in the interpretation process by developing a flexible statistical framework that is capable of modeling the mean trajectory and not depending on assumption or sample size too much. Such a framework can be useful to integrate the information from limited data and provide insights for experimental designs.

In Chapter 5 we develop a flexible statistical framework for production trajectories. In biological experiments, recombinant molecules are produced over time under various experimental conditions. The data generated are typically longitudinal, and comparing optimal trajectories across different conditions is a major topic in longitudinal data analysis. While the expected amount of the ingredient is usually the object of interest, obtaining multiple replicates for a comprehensive ANOVA approach can be challenging in real-life experiments due to resource constraints. In many cases, the focus is not on the level of the target molecule itself, but on some nonlinear functionals of the production trajectory, such as the time to reach a specific value, the maximum production level, or the maximum productivity. Analyzing protein production trajectories across different experimental conditions poses statistical challenges. Monotonicity of production trajectories is necessary to properly define quantities of interest, but limited data points require information to be borrowed across trajectories to compare parameters. Statistical inference procedures relying on large sample theory may be inaccurate due to limited data and monotonicity restrictions. Additionally, nonlinear parameters and process metrics of interest prevent the use of the standard ANOVA framework based on linear model theory. To address this, we develop a flexible ANOVA framework by representing trajectories in B-spline bases, incorporating monotonicity constraints through linear inequality constraints, and fitting trajectories through a constrained least squares regression procedure using quadratic programming. For statistical inference, we use bootstrap or resampling procedures, comparing the efficacies of residual, parametric, and nonparametric bootstraps to resolve the practical issue of limited data. To control the false discovery rate, we adopt a technique for simultaneous confidence intervals involving many parameters.

In Chapter 6, we extend the simplified ANOVA framework proposed in Chapter 5 to a generalized two-factor ANOVA framework, which is constructed by incorporating the treatment, time effects and the interaction between them. It is able to leverage the impact of different treatments and interactions of both treatment and time and, more informatively, how different they are across treatments. These are also key experimental studies of interest. It quantifies the treatment effects and interaction effects between time and treatment to be studied efficiently. We demonstrate the efficiency in the estimation of the coefficients that can still be solved by quadratic programming. Moreover, we illustrate the capability of this framework in modeling the correlated longitudinal data and making inferences under limited data settings. Such extension yields insightful variabilities coming from different sources. We show in simulation studies that the extended model is able to yield an effective fit and useful description of the data variabilities, with limited data.

CHAPTER 2

Color-complexity enabled exhaustive color-dots identification and spatial patterns testing in images

2.1. Introduction

Spray technologies via unmanned aerial vehicle (UAV) for liquid chemicals of fertilizers, herbicides and pesticides are at the stage of intensive research and developments [**31**]. From economic and environmental perspectives, these technologies are deemed vital in Precision Agriculture [**78**]. Since its wide uses will not only save costs from many aspects, particularly on human labor and illness, but also add capabilities of dynamic and optimal management. However, the success of such technologies heavily relies on effective evaluations of their performances in terms of efficiency and precision. There might be many ways of making such evaluations. One fundamental way is to evaluate their performances by testing whether the sprayed liquid droplets are relatively homogeneous in size and distributed in a spatially uniform fashion upon a target area.

Recently it has become very common that companies and research labs design their own experiments to facilitate such fundamental testing. One key step of this testing involves color image analysis consisting of two coupled computational tasks: exhaustive color-dots identification and spatial patterns extracting and testing. These two tasks are in accord with two natural perspectives of color complexity of a color image [**109**]. It is crucial to be able to exhaustively identify all targeted color dots of all sizes on a target area. Since each dot of sprayed chemical gives rise to two pieces of information: its amount of chemical and spatial location. Exhaustive search and extraction often are difficult to achieve computationally. Even though color identification is a major topic in computer science, those publicly available techniques, such as Contour on grayscale and other color segmentation techniques through computer package like OpenCV, are not optimal, nor practical choices for dealing with heterogeneous shading on color images. In order to better perceive and appreciate the color complexity and its induced computational challenges facing us in this study, some relevant information of the physical nature of color is essential as well as necessary.

Humans recognize different colors when visible lights are received by photoreceptors. Those colors human perceive are indeed grouped into categories, but due to other processes and not only photoreceptors' response. In the physical world, the region of the wavelength of the rainbow of visible lights is between 780nm to 380nm (in decreasing order at nanometer scale). That is, Red has the longest wavelength (around 700nm) and the smallest frequency (428 Terahertz (THz)), while Purple has the shortest wavelength (380nm) and the largest frequency (714THz). The Yellow is in between with wavelengths around 580nm and frequency around 517THz. Such linear ordering of wavelength and frequency of visible light is turned into a circular order of colors, called the color wheel, through human perception. Isaac Newton has studied this nonlinearity in the 17th century. Now we know that the three types of cone cells in our human eyes, which specifically respond to three visible lights: Red, Green and Blue, collectively generate all colors shown on a color wheel [**51**, **95**, **106**].

How the lights come from the outside world into our eyes or lens of cameras is the topic of the physical nature of color. Outside of the eyes and camera, the color should be described exactly through the spectral reflectance within the wavelength region from 380 nm to 780 nm. This spectral reflectance and diffuse reflection [**43**, **55**], depend on the nature of the material and its surface properties, light source, viewing angles, observer (the properties of eyes or camera for imaging) and other surrounding objects. Light not only reflects from the surface of material, but also penetrates beneath the surface and then scatters. Blue has the highest scattering intensity than that Red and Green due to its short wavelength and high frequency. How visible lights are composed into colors inside eyes and cameras is the topic of human trichromacy. The RGB color space is resulted from technical convolution of three monochromatic spectral stimuli: R(λ), G(λ), B(λ), curves [**51**, **95**, **106**]. This RGB color model is mutually transformable with the HSV model: Hue, Saturation and Value [**107**]. HSV and its variant models were designed and are popularly used by computer graphics researchers [**56**].

Besides RGB and HSV models, there are color systems being used with a focus on various color characteristics in different topics and industries. As for color-printing techniques, two color-systems are popularly used for different purposes. The CMYK (Cyan (close to blue), Magenta (close to red), Yellow and Black) system is primarily used in a printing factory, the PMS (Pantone Matching System) system is designed to identify exact color needed, such as in a paint shop. This study is in the opposite direction. We try to avoid or at least limit involvement of human visualizations.

In contrast with RGB, which was defined and adopted in 1930, CIELAB color system: L*a*b*, was defined by the same International Commission of Illumination in 1976 [**1**]. Here L*is referred to as lightness: white-vs-black. a* is indexed relative to green-vs-red, while b* is indexed relative to blue-vs-yellow. Its original intention is to approximate human perception. To convert RGB or CMYK coordinates to or from L*a*b* is not straightforward. We must know the reference Illuminant of RGB and CMYK beforehand. This requirement is not practical for images taken in the field study. It is well known that CIELAB lacks perceptual uniformity, particularly around the blue hues. This character of CIELAB surely raises a special concern, especially for using it as a color system in this study. Therefore we mainly focus on RGB and HSV in this paper.

With respect to a chosen color model or system, a color image is indeed a large data matrix. Theoretically, any data set has a Kolmogorov complexity in Information Theory. This complexity is referred to as the shortest length of a computer program that can regenerate the data as one whole [**70**]. Though this is not a computable concept, it is highly relevant to the majority of data analysis involving finding its pattern-based information content, such as in this study.

Unlike the concept of Kolmogorov's complexity, the color complexity is simply considered through human's data visualization, not computers' data generation. This is a concept not yet being well established in Color Theory. A colored image is a huge dataset because of its many thousands or millions of pixels, and each pixel is coded with 3D RGB and HSV coordinates. The color complexity of an image can be perceived and computed from two perspectives: color variety and color distribution [**109**].

In this paper, both perspectives are being studied across five experimental images. The color variety is referred to how many "distinct" colors are present within an image, while color distribution is referred to 2D spatial distributions of distinct colors on the 2D Euclidean plane of the image. We discuss these two perspectives computationally in this study. Even though, each of the five images seemingly has only two major colors by experimental design, the color variety within the RGB space surely would be much more than 2. There are many underlying factors that could affect the large and fine scales coloring of an image. We just list three key factors here as follows.

First, the mechanism of light's reflection upon an unsmooth and uneven colored paper media involves with complicated physical nature of color, as been briefly depicted with some basic facts above. Secondly, colored dots are of many sizes, and indeed could contain heterogeneous color-pigment intensities. Thirdly, swiftly varying lighting conditions could go through the lens of cameras when the images were taken. Due to these reasons and beyond, we expect that one image would give rise to a RGB point-cloud occupying many 3D coordinates within the RGB space. We can intuitively define an image's color variety as the relative size of its RGB point-cloud on the finest scale with respect to the size 256^3 of 3D RGB space $[0, 255]^3$.

In this study, we analyze five images collected by the researchers from GEOSTA & Technology Inc., which conducted the experiments of spraying liquids via drone. The experimental setting realistically mimicked the mechanism undertaken by an unmanned aerial vehicle (UAV) when spraying liquid chemicals on the fields. Multiple yellow Teejet water-sensitive papers (part# 20301-1N) are laid on flat ground when an UAV machine was employed at each trail. Such yellow paper shows stained blue-purplish colors wherever being exposed to aqueous spray droplets. The five testing papers that are analyzed in this paper are collected from five different spraying trails with distinct mechanical parameters that control spraying mechanisms. Water is used to do the test and there was no pesticide involved. More information on the design of the UAV machine and nozzle's spray angle-vs-coverage relational information is contained in [**98**].

After spraying, each test paper was then individually placed on a specific background-setting. Its camera image was taken. The five background-settings are slightly different from each other. Therefore, the five camera images of the five testing papers were taken under different lighting conditions. For instance, see Fig 2.1A for one experimental image. Then, each of the five images is individually converted into one RGB and one HSV data sets, see Fig 2.1B for both color models. Within these five pairs of data sets, the joint characteristics of spatial geometries and size distributions of blue-purplish colored dots are used to evaluate the effectiveness of the five experimental mechanisms.

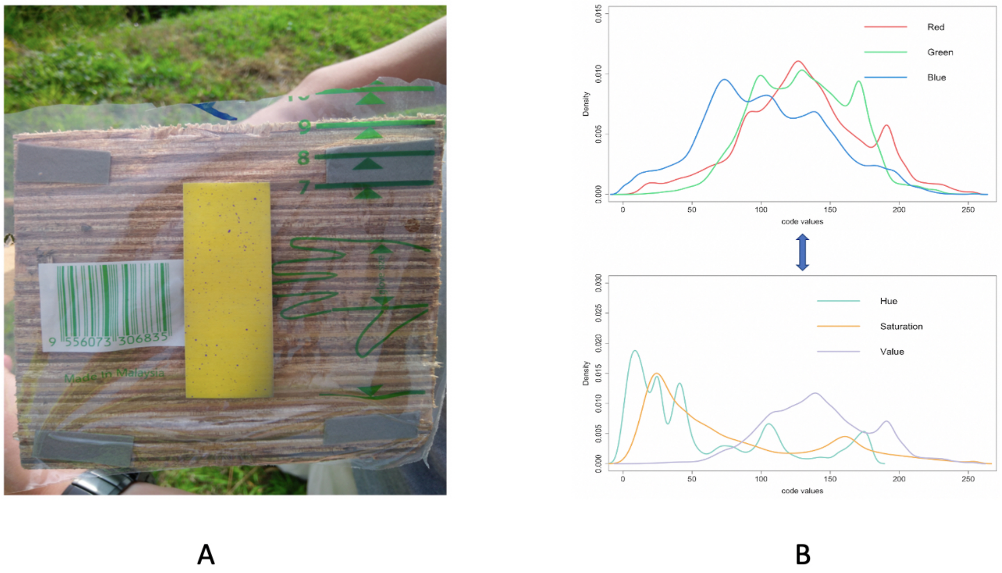


Figure 2.1. Original image and RGB density curves. A: Original image with purple dots in various sizes and shapes located on the yellow test paper; B: RGB model and HSV model.

Thus, the goals of this study are to be able to computationally extract precise spatial and distributional information of all blue-purplish dots and to prescribe their interacting relations as joint characteristics. To achieve these goals, blue-purplish color identification is the first computational technique we would develop in this paper. This technique needs the capability of capturing color-dots of all sizes with varying shapes. Based on RGB and HSV color models, we expect that each of the five images would have a rather small color complexity. It is because that the two major colors would make three RGB coordinates as three features of 3D data points highly associated. That is, all RGB point-clouds are to have a two-hill shape occupying small volumes in the RGB space.

Various algorithms and techniques such as thresholding techniques [**67**] and clustering, have been developed for image or color segmentation [**53**]. Most of them can be implemented through OpenCV. Applying color thresholding typically requires human involvement in color identification by looking for spatial spots in an image that match a fixed range of colors. Such a range of colors is either pre-specified or being taken out by humans from an image. Such a priori kind of knowledge makes the most direct and clear distinction between existing popular color-Identification packages, such as OpenCV, and goals of color Identification in this study. Since this range of color within an image is one specific information to be computed and learned from the image itself. That is exactly one of the challenges facing us in this study. Since the heterogeneity in targeted color created by varying shades and tones makes specifying such a precise range of colors extremely difficult.

Specifically, shades and tones affect colors locally, while lighting conditions would impact colors globally. Such lighting conditions are nearly uncontrollable for any camera image taken outdoor under the Sun. Such global and local factors make color distributions vary drastically from one image to another. This varying nature does not fit well with many color-identification techniques based on computational methodologies, such as K-means [**37**] and variants of thresholding techniques [**2**], that require prior knowledge of object and background distributions. Further, irregular shapes are another characteristic of the color dot in our testing papers. Such irregularity in shape and varying in sizes together with a large number of dots make the majority of segmentation techniques [**53**] nearly useless. The widely spreading tiny color dots would act like noise that seriously compromise the effectiveness of such segmentation techniques.

The second computational technique is to characterize spatial geometries with respect to varying dot-sizes as the targeted joint characteristics. As for the perspective of color dot distribution, we only focus on the targeted purple color's distribution because the Yellow color is the background color. Since the spraying mechanism is a mechanical one. So, we expect the potentials that the purple color's distribution might be far away from uniform, particularly for dots of large size with high intensity of purple. Due to the largeness of the number of purple color-dots, we focus on spatial uniformness in the sense of density of purple dots, not their spatial coordinates. In order to bring out the sense of density, we divide the test paper into a feasible number (400) of squares, and categorize their densities in terms of their distributions of categorical sizes of purple dots contained in them. For one cumulative density category, from the largest to smallest, we propose to build a minimum spanning tree (MST) to connect squares. A MST is a rather flexible spatial structure. Its characteristic distribution of the distance between immediate neighboring-squares would be the basis for testing spatial uniformness.

By simulating MST under spatial uniformness assumption upon all squares, we compare the observed MST with all simulated MSTs. To facilitate such a comparison, we extract one MST-based distribution of the distance of neighboring squares from each MST. Further, we transform each distance distribution into a histogram with common data-driven bin-boundaries, and then collecting all vectors of proportions into a matrix. We build a Hierarchical Clustering (HC) tree among these distribution-IDs. Then, we develop a new algorithm to calculate p-value based on the binary structure of HC tree-geometry. We then repeatedly perform the same testing on uniformness by including less dense squares in a cumulative manner. This p-value computation is somehow novel in the sense that it is calculated based on a series of odds-ratios along a descending tree-path leading to the observed MST-tree-leaf. Such HC-tree-based spatial-uniformness testing and evaluating the reliability of the testing results are performed with respect to each of the multiple cumulative density scales. We believe that such a multiscale spatial 2D uniformness testing seemingly offers a relatively new perspective of spatial data analysis.

In comparison, our goals and computational developments for evaluating the effectiveness of an UAV's spraying mechanisms in this study are rather unique comparing with existing color-identification techniques. For instance, in medical image analysis, the focus is placed on identifying color-dots under a rather limited and well-controlled range of lighting conditions. Therefore, the issues of shade and tone are not as serious as in the open field settings. Further, color images are highly sensitive to environmental and operational conditions, such as lighting and shadowing under weather and operations. It is realistic, practical and even necessary to extract designated color pixels with respect to data-driven RGB ranges, not fixed ones as used in popular color identification approaches, such as the aforementioned color thresholding and others. Since the effects of shade and tone vary from one color image to another color image, we need robust computational efforts to accommodate such varying effects.

2.2. Method

The original dataset is in the format of an image, with a dimension of 4608×3456. We use R programming to read and load the image, transforming it into a 3-dimensional array. Each dimension corresponds to a 4608×3456 matrix with one color channel as its entry. We further reconstruct the data as a large matrix with x-, y-axes, and three color- channels as columns. We make use of the array-like object to reproduce an image and illustrate our results, and a matrix for computation.

Each image gives rise to two RGB and HSV data formats, as shown in Fig 2.1B. They are mutually convertible. We remove most of the background and only focus on the area containing the yellow test paper, which contains around 10^6 pixels. The image of the test paper can be reconstructed as any one of RGB or HSV $10^6 \times 3$ matrices. That is, from perspectives of pixel-wise 3D intensities of RGB and HSV, a color image data here is in the form of structured data. Each pixel-specific color data point in an image like the one in Fig 2.1A.) are referred to two 3D color measurements: RGB ($[0, 255) \times [0, 255) \times [0, 255)$) and HSV ($[0, 255) \times [0, 255) \times [0, 255)$).

Before our developments in this section, we mention the fact that the five images are likely to subject to mixing lighting conditions. Such conditions can cause uniform shading effects across the entire image. Such uniform effects can be dealt with by taking off via various imaging processing techniques. Nonetheless, we do not perform such a data pre-processing step. Since we see that the shading effects seen and considered in this paper are primarily local and heterogeneous. They are mainly caused by unevenness of the test paper, or some surrounding objects, such as human hands holding the test paper, that indeed block lights from shining onto small parts of the test paper when it was posted for its images to be taken. Such characteristics of locality and heterogeneity on an image are different from the uniform ones.

Our machine learning-based approach is proposed to identify localities affected by the shading effects and to figure out the varying degrees of the effects in a divide-and-conquer fashion. Further, such local and heterogeneous characteristics of shading effects vary from one image to another image. Therefore, we need at least a learning region being anchored for each image to train our computing methodologies on typical areas, and then modify our learned methodologies one way or the other to adapt to shading effects upon affected localities. We believe that the local and heterogeneous nature of shading effects, as being present across all five images studied here, are not uncommon in real-world applications.

We propose computational algorithms to resolve this color identification issue. One physical fact of color theory plays an important role: the three dimensions of RGB or HSV are highly associated. This fact interestingly and very important points to that, among the 256^3 unit cubes (of size $1 \times 1 \times 1$), a color image's millions of pixels typically only collectively occupy a very small number of cubes. That is, a natural color image usually has a very small "color-complexity".

The color-complexity of the test paper in Fig 2.2A is 0.002. In contrast, the color-complexity of the whole image in Fig 2.1A is calculated as $\frac{393014}{256^3} \approx 0.023$. That is, the color-complexity of this test paper is only one-tenth of the original image. Therefore, we indeed deal with only several thousand, not a million, of distinct colors. This is the underlying reason why our computing cost is low and our color identification is effective.

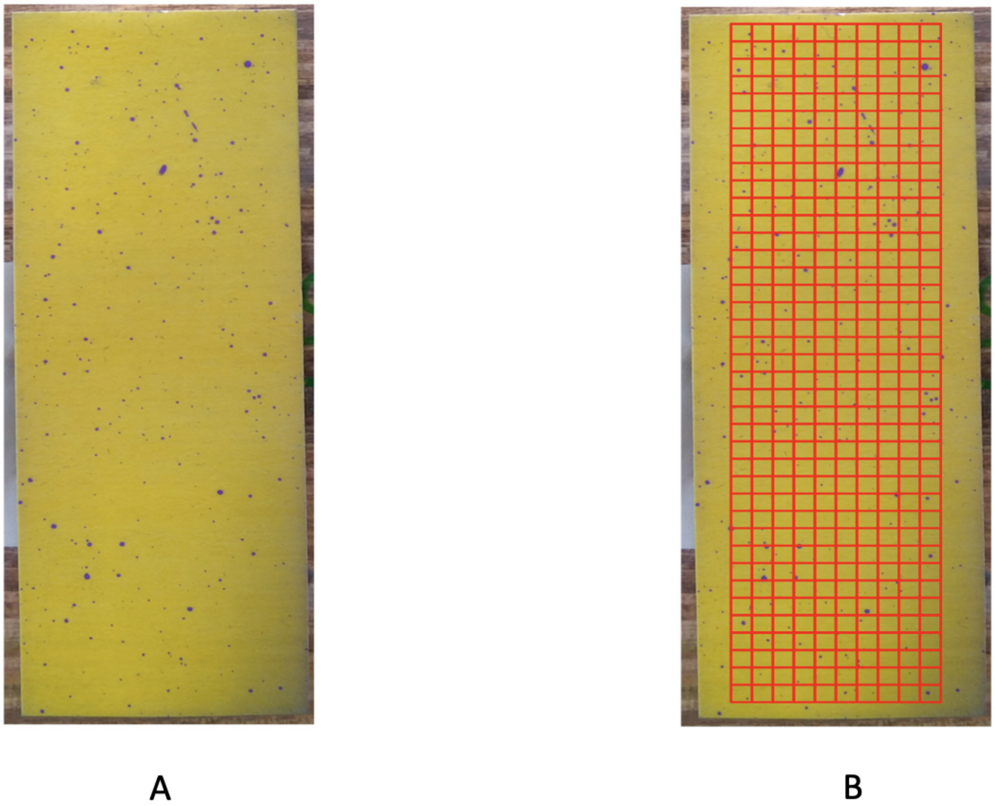


Figure 2.2. A: reduced image; B: focal area.

Upon one test paper (yellow colored-strip), under either RGB or HSV color models, our first development is aimed at exhaustively identifying all "purple" dots of all sizes. Here color "purple" is meant to be an unspecified 3D region within the 3D discrete space $[0, 255)^3$ that reveals purple color through our visual system. After nearly exhaustively extracting all purple pixels, we build purple dots as a connected network based on a common choice of the neighborhood on the 2D lattice. Then, we measure each dot's size, and classify them into several size categories.

After collecting almost all purple dots and sorting out their multi-scales size-categories, two key difficulties facing us here are: the largeness of the number of color-dots and geometric representations of color-dots. These two difficult aspects correspond to two kinds of uniformness: dot-size and spatial. Upon dot-size uniformness, we aim to figure out whether the spraying machine's mechanical design is proper or not. We particularly pay attention to the behavior of the right tail of the dot-size distribution (large and very large ones). While, upon the spatial uniformness, we need a practical unit that can embrace effectively the concept of spatial density of dot-locations. We also need a simple enough geometric representation to embed all involving units, so that structural information of spatial distribution can be extracted. Such computational endeavors for multi-scale spatial pattern extractions and testing spatial uniformness upon all size-scales are computationally developed in this study.

In this paper, we develop computing algorithms to resolve such coupled computational tasks. A block of diagram clarifying our proposed method process is displayed in Fig 2.3. We apply our algorithms to five experimental images under rather distinct lighting conditions. We exclusively use one image for illustrating our computational developments (Fig 2.1A), followed by the results of the rest of the 4 images.

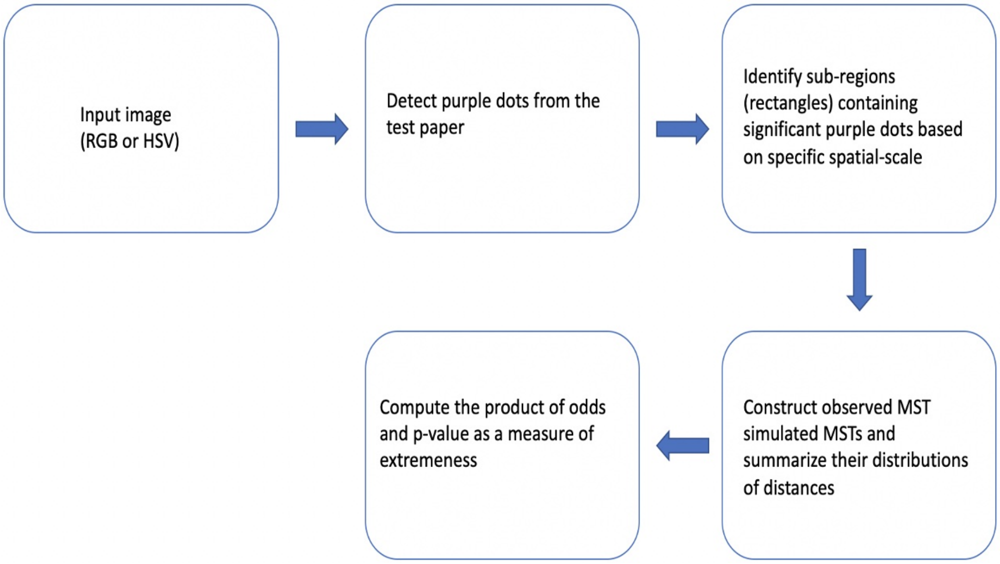

Input image
(RGB or HSV)
Detect purple dots from the test paper
Identify sub-regions (rectangles) containing significant purple dots based on specific spatial-scale
Construct observed MST simulated MSTs and summarize their distributions of distances
Compute the product of odds and p-value as a measure of extremeness

Figure 2.3. A block of diagram of the process of our method.

In order to exhaustively identify all pixels of a human-designated color, under the RGB system, we apply the Hierarchical Clustering (HC) algorithm as an unsupervised machine learning approach in a divide-and-conquer fashion due to the largeness of pixel numbers. The tree-leaves of the designated color-branches are recovered with their spatial coordinates. The effectiveness of this color-identification can be inspected via zoom-in and zoom-out visual validations based on local as well as global geometries. Multiscale inspections conducted to make sure our algorithms' validity and efficiency are necessary due to the huge number of pixels.

Further, since the collection of identified irregular shapes and size color dots is also large in number, we divide the whole spatial region into rectangles of the same size and extract their categorical features of color-dot-density. Then two brand-new computational approaches are devised for our data analysis. First, we propose to construct one minimum spanning tree (MST) upon one cumulative scale of color-dot-density to capture spatial characteristics. Secondly, we develop an odds-ratio-based algorithm to compute a p-value from a clustering-tree, which is built upon a set of summarizing vectors extract from the observed MST and many simulated MSTs under the uniformness hypothesis.

2.2.1. Identification of purple pixels. As shown in Fig 2.1A and Fig 2.2A, it is clear that this test paper contains two main color families, yellow and purple, among the one million (10^6) pixels. It is also evident that it contains areas of heterogeneous intensities of shades across the entire test paper. The presence of such data complications is rather common in the majority of real-world color images, It becomes part of the nature of data from Precision Agriculture. Since images might be taken under drastically distinct lighting conditions: with or without sun lights across different parts of days. Further, it is well known that the human's visual system via the brain and eye is subject to color illusions. Such an illusion makes us identify the same object with different colors under shadows as well as different backgrounds. Thus, any heterogeneously shaded image, in general, poses various challenges on color identification. One of the challenges is: How to do color identification in a data-driven fashion? In other words, it is a necessary capability of identifying color in any image from the perspective of a computer, not a human.

To make computing feasible via computer, we need to have an idea of how many distinct colors are indeed contained in the test paper. This is the concept referred to as "color-complexity". Given the discrete nature of color data, it is crucial to ask: how many unit cube of $1 \times 1 \times 1$ among the 256^3 "color-unit-cubes" are indeed occupied by the one million color-pixels in the test paper? The answer is 28126. So the color-complexity is only $\frac{28126}{256^3} \approx 0.002$. If we enlarge the scale of the unit cube to a scale of $10 \times 10 \times 10$ cube, we checked and found that all potential colors contained within such a cube are still rather "uniform" to our raw eyesight. And, with respect to all pixels in the test paper, there are 880 among $26 \times 26 \times 26$ of such cubes being occupied. The color-complex of this test paper on this larger scale is $\frac{880}{26^3} \approx 0.05$. Hence, we decide to begin our machine learning computations upon this scale first and go back to the unit scale afterward. It is worth emphasizing that such low color-complexity is made possible by very high non-linear associations among R&G&B and H&S&V. This is the underlying foundation to build data-driven algorithms for color identification.

Then we build a geometry among these 880 uniform color cubes. This geometry is intended to serve as a platform for our color identification. We choose this geometry to a tree for computational simplicity and practical applicability. We construct hierarchical clustering(HC) trees as follows. We use the center of mass (3D average) of pixels contained in such a cube as the cube's representative. Upon this collective of 880 representatives, the HC algorithm can work efficiently.

For completing our protocol of color identification, we take a step to tentatively avoid shady areas and background noises. Even though only involving a minority of pixels, their inclusion could yield non-negligible errors. To this end, we choose a rectangle area within the test paper as our "focal area", as shown in Fig 2.2B. This focal area is divided into 39 rows. Each row contains 2.5×10^3 pixels (Fig 2.2B), and is further divided into 10 squares.

Our color identification begins with the following row-by-row operation. For each one row's 2.5×10^3 pixels, we identify which color-cube it belongs to and then find its color-cube representative. The resultant set of distinct color-cube representatives has a size smaller than 880, surely is much smaller than 2.5×10^3. Upon this row-specific set of color-cube representatives, we apply the HC algorithm. For each row-specific HC-tree, via its bifurcation, we collect the representatives within the smaller branch as being designated as "purple" ones, while those in the larger branch as being "yellow". We further use ROC curve analysis for validation checking to avoid misclassification due to uncontrolled environmental and lighting conditions. This validation check is performed upon each square within each row.

The ensemble of color-identification on the focal area via the RGB data file is shown in Fig 2.4 together with results of the square-by-square validation check. There are three squares that have obvious misclassification. We "clean" these three squares by assigning all pixels in these three squares into the yellow group.

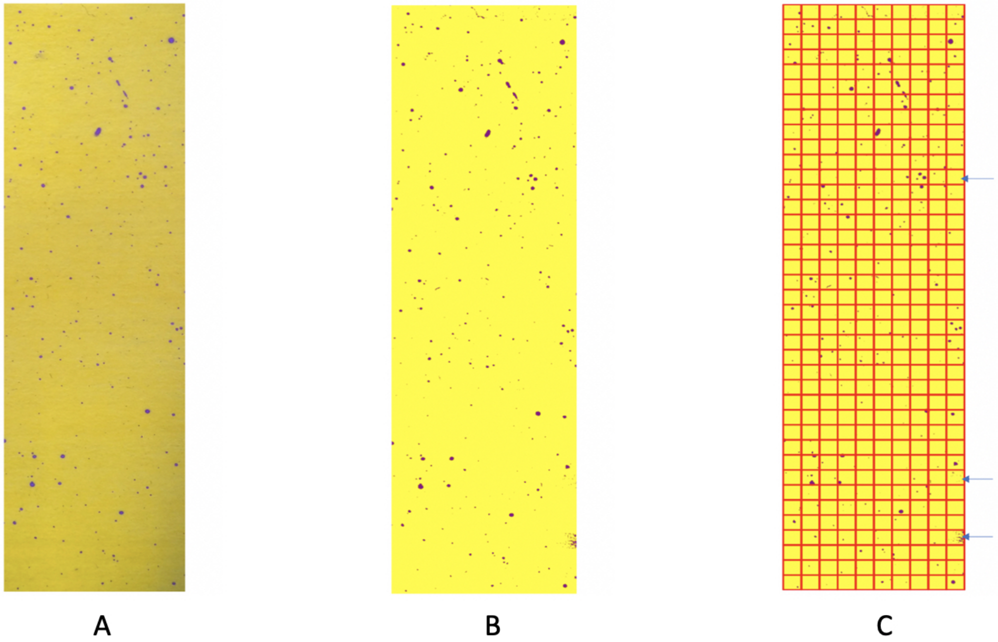


Figure 2.4. Color-identification on the focal area A: original focal area; B: predicted focal area by RGB file; C: validation check indicating three squares need "cleaning".

2.2.2. Comparing with existing approaches. As aforementioned, one test paper likely includes regions under varying shading conditions due to the photographing condition and experimental setup when the image was taken. Heterogeneous shaded images are vividly seen as well in the four images shown and analyzed later. Such existential shading will complicate choices of grayscale, and consequently, reduce the efficiency of the Contour approach significantly. We apply these aforementioned methods and algorithms to one part of our image, as comparisons to our proposed method.

The segmentation results are shown in Fig 2.5 that indicates one example of such local impacts of the shading effect. It is observed that the shading effect exists along the right bound, especially at the bottom right. All three methods are sensitive to noise and shading, and thus fail to correctly segment the purple objects and yellow background. The poor identification by K-Means and the Mean thresholding method are affected due to the shading. As for the color thresholding method, we set the RGB code as (247,191,252) as the upper bound (light purple) while (75,0,130) as the lower bound (dark purple). Since so many dots of varying sizes look purple, it is not possible to specify a 3D color region to cover all purple dots.

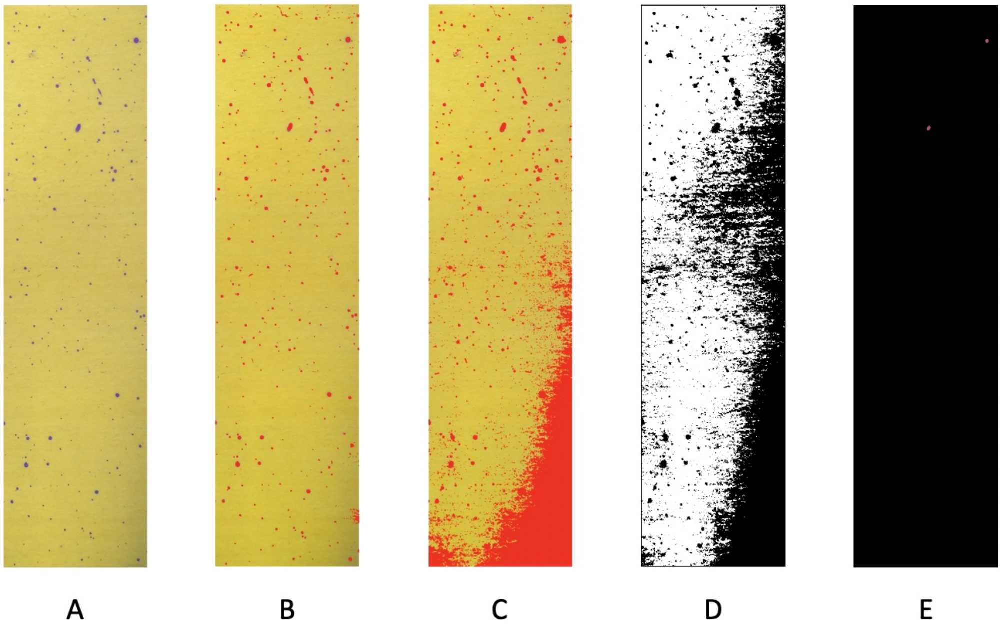


Figure 2.5. Color segmentation using different algorithms and techniques. A: part of the original image; B: our proposed method; C: K-Means; D: Mean thresholding; E: Color thresholding.

We, therefore, conclude that the color thresholding technique via OpenCV is neither feasible nor practical for the exhaustive search purpose in our case. Hence, due to practical issues caused by shading and lighting effects, such classical applications are less effective, They neglect key color structures contained in the image. Thus, an effective strategy as proposed in the above subsection becomes necessary.

2.2.3. Recovering the whole yellow test paper. Given that pixels outside of the focal area have a higher potential for being subject to shade or other noises, their color identification needs extra effort. We propose the following remedy based on our experiences derived from our explorations and experiments. Upon RGB data format, we need to employ $1 \times 1 \times 1$ small RGB color-cubes, denoted as the scale of "$n = 1$". That is, we need to drastically sharpen the color-uniformness within each color-cube. So we have to pay more computing cost to achieve the goal of color identification with RGB data, even though we still enjoy the reduction of color complexity because only 0.2% of $1 \times 1 \times 1$ RGB unit color-cubes are occupied.

Consequently, we collect the centers of all color-cubes, which have ever been occupied by an identified purple pixel in the focal area. And likewise collect the centers of all color-cubes, which have ever been occupied by an identified yellow pixel in the focal area. Among the pool of these two collections of centers of color-cubes, we compute the closest neighbor to each pixel outside of the focal area and then declare the color-identification accordingly. In this way, we are able to capture the majority of purple pixels and avoid misclassification as much as possible.

As for HSV data format, we still employ the scale $n = 10$, i.e. $10 \times 10 \times 10$ HSV color-cubes. We obtain the recovering by both the RGB file and HSV file separately. It turns out that the RGB file helps identify more pixels in smaller purple dots, while HSV file helps identify more pixels near the bottom and top where the RGB file fails. The two results suggest a better recovering scheme as simply combining these two results together. All results are shown in Fig 2.6.

2.3. Testing uniformness via sizes

As it is intuitively known that a spraying device typically mixes air with liquid, and then pushes the mixture out. The mixing of air and liquid is determined by a set of tuning parameters. Mechanically speaking, different sets of tuning parameters surely give rise to distinct degrees of inhomogeneous mixing. Consequently, the droplets out of the device are likely heterogeneous in size. So, some tuning parameters are better than others. One merit of exhaustive identification of targeted color-dots contained in an image is to check the validity of a parameter-setup of the spraying device. For this merit, there are two natural measure sizes of a droplet, which is an identified connected purple-pixel. The first measure is to count the number of connecting pixels.

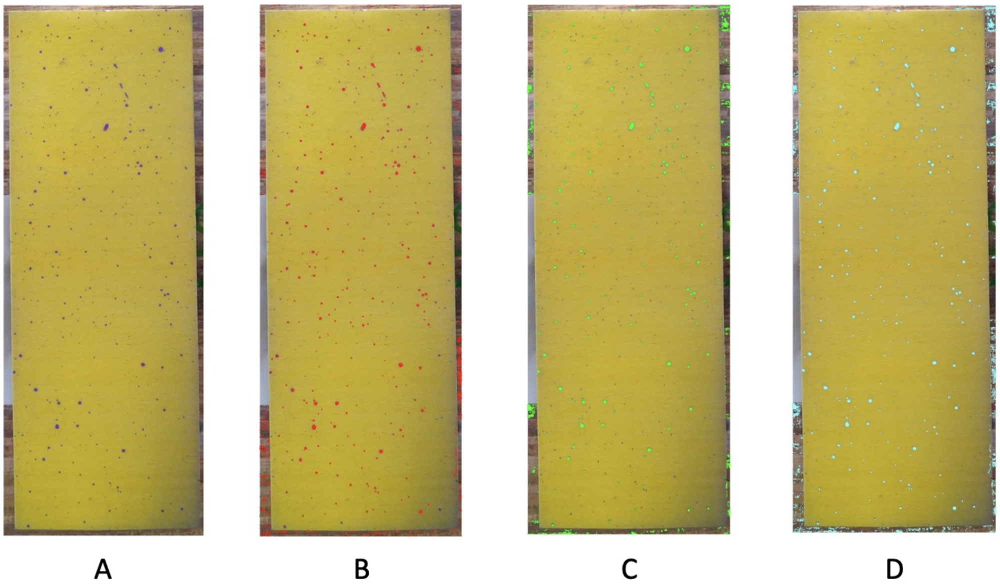


FIGURE 2.6. Performance by using different files A: original reduced image; B: recovering (in red) by RGB file ($n = 1$); C: recovering (in green) by HSV file ($n = 10$); D: recovering (in light blue) by combining RGB file ($n = 1$) and HSV file ($n = 10$)

The second one is the radius of the smallest circle containing all connecting pixels. Accordingly, the best set of tuning parameters should ideally produce the Poisson distribution with respect to the counting measure, and an Exponential with respect to the continuous measure.

We consider the target collection of color-dots identified via the approach of combining the RGB and HSV data, see Fig 2.6D. We first compute the MLEs of intensity parameters, λ_P, and λ_E, under the Poisson and Exponential distribution assumptions, respectively, based on the two data sets derived from the target collection of purple-dots within the test paper.

Based on the pixel-count data set, the Poisson distribution specified by MLE of λ_P is computed and superimposed onto the histogram constructed based on pixel-counts from the target collection of purple dots, as shown in Fig 2.7A. It is evident that many identified purple-dots have large pixel counts that can not be accounted for by Poisson distribution. We can draw a similar conclusion based on the dot-size distribution with superimposed Exponential distribution specified by MLE of λ_E shown in Fig 2.7B.

While the Q-Q plot Exponential distribution specified by MLE of λ_E is compared with empirical Q-Q plot of continuous purple dots sizes, as shown in Fig 2.7C. We see evident departures from this

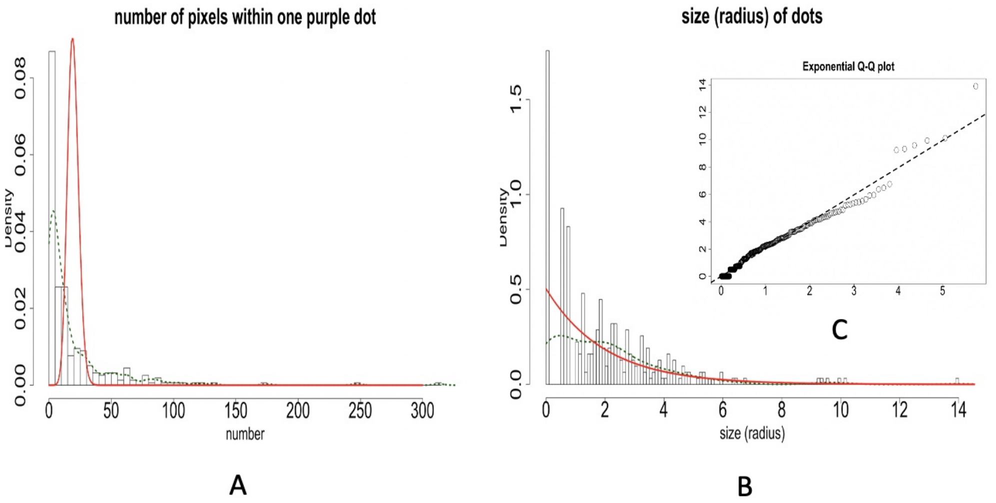


FIGURE 2.7. Distributions of pixel-counts and dot-sizes A: The histogram of pixel-counts of identified purple-dots superimposed with Poisson distribution specified MLE of λ_P(red curve) and kernel density estimates (green dash curve); B: The histogram of dot-sizes of identified purple-dots superimposed with Exponential distribution specified MLE of λ_E (red curve) and kernel density estimates (green dash curve); C: Empirical Q-Q plot (Circle curve) vs Exponential Q-Q plot specified by MLE of λ_E(dash line).

Q-Q plot comparison. Further, we run the Kolmogorov-Smirnov test, which suggests the observed dot-size is not following an Exponential distribution (with p-value < 0.05)

2.4. Testing spatial uniformness via rectangle neighborhood

In this section, we construct our major algorithmic developments for testing against the 2D spatial uniformness. We adopt the concept of the 2D neighborhood into 2D spatial characteristics. The reasons behind this are that the number of identified purple-dots is too big, and their sizes are rather heterogeneous. This neighborhood concept directly links to the idea of spatial density, which is a proper expression for addressing spatial uniformness here.

Given that we specifically divide the entire target area into 400 small rectangles, one rectangle is taken as one 2D neighborhood. On this collective of rectangles, we pretend as if they are uniformly colored with an intensity (density) of purple depending on all purple-dots contained in it. In this fashion, we consider the 2D spatial uniformness among 2D-entities of 400 rectangles. In addition, the radius of the smallest circle containing all connecting pixels is regarded as the size of a purple dot and thus all identified purple-dots will be classified into three categories of sizes: small, medium and large, according to their sizes. So intensity of purple dots in a rectangle would be also categorized, as given below. We apply the Hierarchical Clustering (HC) algorithm to guide this categorization. The categorizing protocol is devised as follows.

We count the numbers of small, medium and large purple-dots contained in a rectangle as the 3 features for this rectangle. That is, each rectangle of many pixels as an unstructured data format is characterized by a 3-dim vector of counts. Via this characterization, we transform a rectangle into a structured data format. We employ a distance measure that is a weighted version of Euclidean distance in R^3. To reflect larger dot-size giving rise to higher purple-color intensity, this weighting scheme is specified with respect to the 3 averaged sizes: small, medium and large, of purple-dots. With this weighted distance measure, we build a 400×400 distance matrix. A HC-tree is computed and reported in Fig 2.8B.

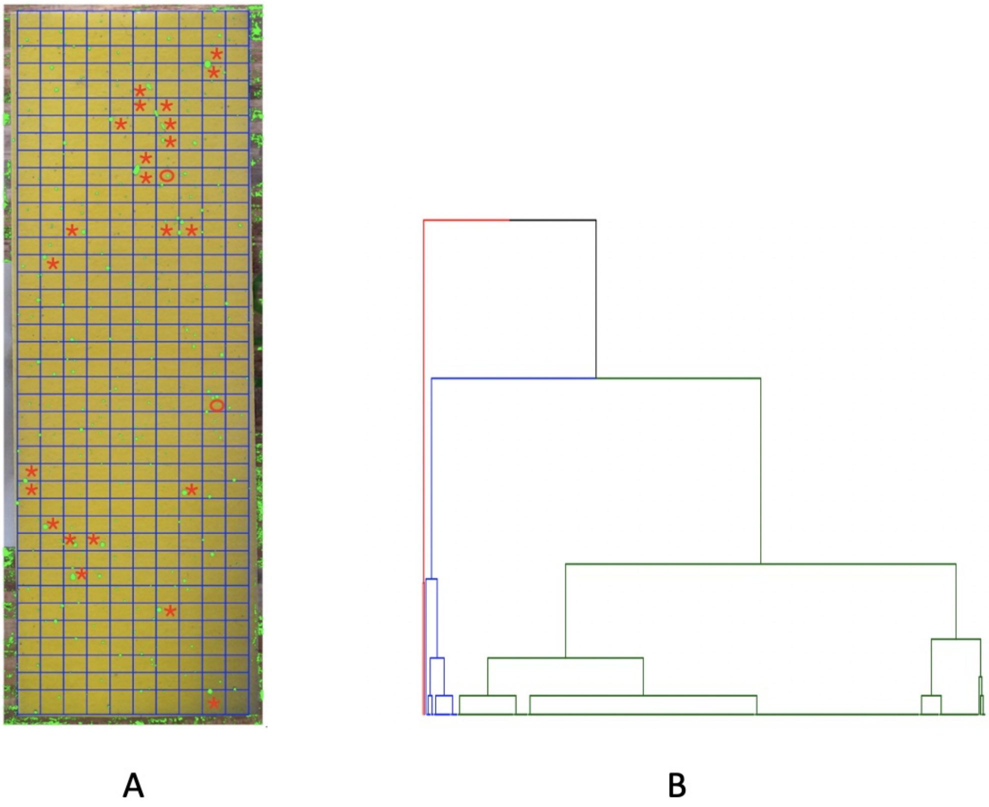


FIGURE 2.8. Significant rectangles are selected by using HC algorithm. A: $\star$ in red: rectangles with ≥ 1 big dot are identified (in blue branch in B); o in red: rectangles with ≥ 2 medium dots (in red branch in B); B: HC tree of 400 rectangles with 3 indicated clusters.

Upon this HC-tree, we can see two small branches (red and blue colored) constituting a clear pattern: their member rectangles either contain at least one large or two medium dots. Locations of these rectangles are shown in Fig 2.8A. This data-driven pattern leads us to explore the intensity spectrum via a Hierarchical Clustering (HC) tree on these 400 rectangles.

More or less based on this HC tree, we further qualitatively determine four categories of density within a rectangle as follows: A rectangle contains

R1:: [Highly-dense] one or more large dots, or two or more medium dots;

R2:: [Dense] one medium dot;

R3:: [Sparse] 2 or more small dots;

R4:: [Extremely-Sparse] only one small dot or empty.

We found that there are 25 rectangles belonging to the [R1] (Highly-dense) category, which are located on the blue and red branches of the HC tree in Fig 2.8B. The spatial geometries of these four categories of rectangles can be seen in Fig 2.9 in a cumulative manner.

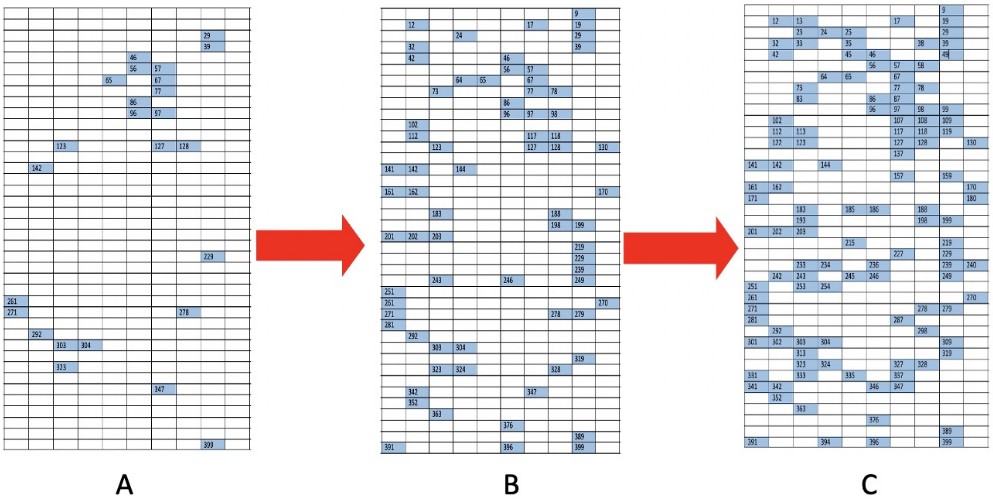


Figure 2.9. The flow of rectangles' 2D-distribution, from large dots to medium and small dots; some specific spatial patterns are present. A: [R1] rectangles; B: [R1-R2] rectangles; C: [R1-R3] rectangles and the blank are [R4] rectangles.

2.4.1. Via Minimum Spanning Tree (MST). Upon these 25 highly-dense rectangles, we construct a Minimum Spanning Tree (MST), denoted as M^{obs}, as shown in Fig 2.10B. The intuitive idea underlying MST is that its tree geometry, which spans a subregion by having one tree-leaf linking to one of its close neighbors, will reflect possibly heterogeneous degrees of spatial concentration among the 25 rectangle members. One way of expressing such heterogeneity in the spatial concentration of a MST is to look through the empirical distribution (or histogram) of distances among all connected immediate-tree-neighbors. Such an empirical distribution (or histogram) is an informative summarizing exhibition for the degree of heterogeneous concentration pertaining to a MST. We particularly lookout for the extremely high concentrations, which will lead a MST's empirical distribution of distance, or its histogram, to reveal a single-mode located at a small distance value.

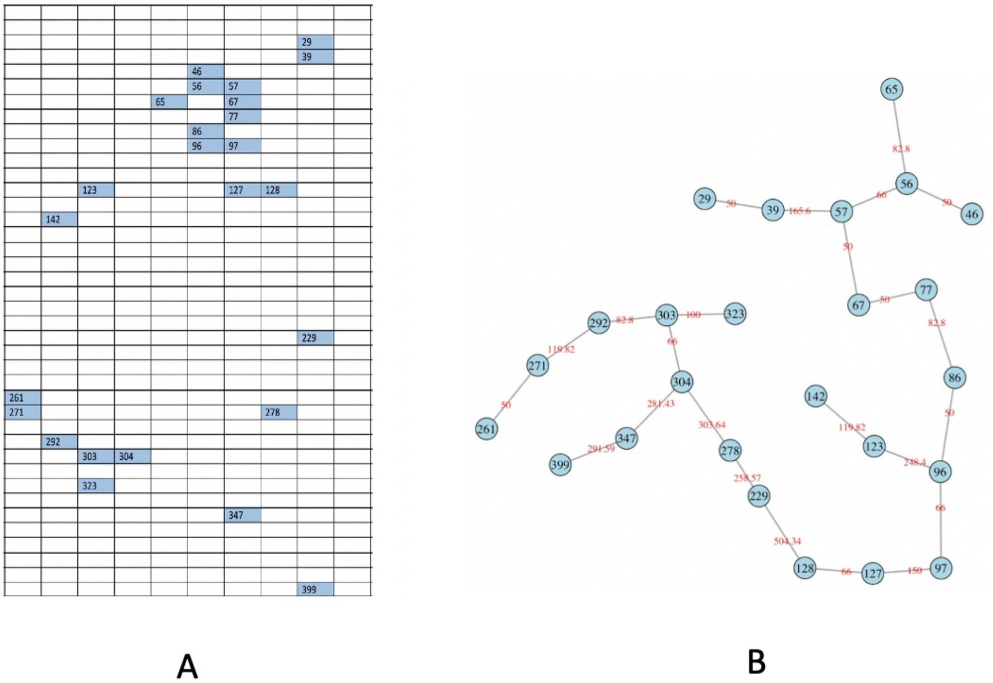


Figure 2.10. Distributions of [R1] (Highly-dense) rectangles and construct MST. A: interested 2D distribution of rectangles with ≥ 2 medium dots; B: using the corresponding Minimum Spanning Tree (MST) to capture the spatial pattern.

With aforementioned focal characteristics in mind, to test whether M^{obs} is coherent with the 2D spatial uniformness hypothesis, we compare M^{obs}'s empirical distribution (or histogram) with $B(= 500)$ randomly generated MSTs' empirical distributions (or histograms). 25 numbers are sampled randomly from a collection of digits $\{1, 2, ..., 400\}$ with equal probability. We repeat this simulation scheme for $B(= 500)$ times with independence. We accordingly generate corresponding B MSTs, denoted as $\{M_b\}_{b=1}^{B}$. So we have B simulated empirical distributions (or histograms) under the spatial uniformness hypotheses. To compare M^{obs} with $\{M_b\}_{b=1}^{B}$ via their empirical distributions (or histograms), we propose two approaches: the Receiver Operation Characteristic (ROC) curve analysis and unsupervised machine learning approach.

2.4.2. ROC curve analysis: The ROC Curve compares a pair of distributions, say $F(.)$ and $G(.)$ via $1-G(F^{-1}(1-w))$ with $F^{-1}(w)$ the quantile function of $F(.)$. So we compute B ROC curves for the B pairs empirical distributions pertaining to B pairs of (M^{obs}, M_b) with $b = 1, ..., B$. So B pieces of area-under-curve (AUC) are calculated. The histogram of B AUC values are shown in Fig 2.11A. The marked 0.5 value of AUC, which corresponds to the case of $F(.) = G(.)$, is seen being far away from this histogram. This fact strongly indicates that the M^{obs}'s empirical distribution (or histogram) is stochastically smaller than the empirical distributions (or histograms) of $\{M_b\}_{b=1}^{B}$ in a persistent manner. Nevertheless, one fundamental drawback rests on the fact that one-dimensional statistics is unlikely to reveal structural differences between two distributions because of their high dimensionality. That is why unsupervised machine learning approaches are needed.

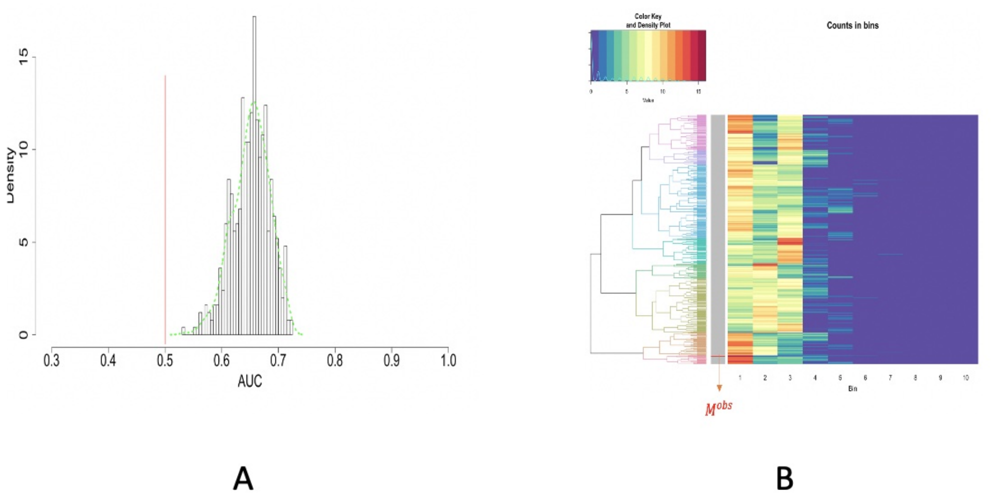


FIGURE 2.11. ROC curve and Unsupervised machine learning approach on testing the spatial uniformness of 25 highly-dense rectangles. A: ROC curve anlysis results; B: HC algorithm with heatmap, product of odds ($PO = 0.023$) and p-value $p(M^{obs}) = 0.002$.

2.4.3. Unsupervised machine learning approach: We want to literally compare these $B+1$ empirical distributions derived from M^{obs} with $\{M_b\}_{b=1}^{B}$. To facilitate such a direct comparison, we pool together all distance values from these $B+1$ empirical distributions and then build a histogram with 10-bins. With such data-driven bin boundaries, we transform each empirical distribution into a 10-dim vector of counts. These $B+1$ 10-dim vectors are arranged along the row-axis a $(B+1) \times 10$ matrix.

Due to the equal total counts on all rows, we simply adopt Euclidean distance and then calculate a $(B+1) \times (B+1)$, with which we apply the Hierarchical Clustering (HC) algorithm to build a HC tree, denoted as $\mathcal{T}$, and superimpose it onto the row-axis of $(B+1) \times 10$ matrix, as shown in Fig 2.11B. The tree-leaf of M^{obs} is marked onto the HC tree upon this rearranged matrix, which is called a heatmap.

The resultant heatmap explicitly shows why M^{obs} is found among an extreme subgroup of the ensemble $\{M_b\}_{b=1}^{B}$. The visible pattern is that the 25 members of M^{obs} have dominantly many extremely small distances among immediate neighbors. This pattern indeed indicates high degrees of concentration among 25 members of M^{obs}. This is a strong piece of evidence against the spatial uniformness assumption. How strong it is? Next, we develop an algorithm to do such an evaluation.

The HC tree $\mathcal{T}$ is binary. Therefore each of $B+1$ tree-leave can be located by a binary tree-descending tracing process. If we adopt a coding scheme to encode the left-branching with a code-0 and right-branching with a code-1 at each internal node of $\mathcal{T}$. Then each tree-leaf is encoded by a binary code sequence. Denote the binary code sequence for M^{obs} as $< d_1^o, d_2^o, ...d_{K_o}^o >$ and a code sequence for M_b as $< d_1^b, d_2^b, ...d_{K_b}^b >$ with $b = 1, .., B$. The coding lengths K_o and $\{K_b\}_{b=1}^{B}$ vary from one tree-leaf to another tree-leaf.

Further, with the binary code sequence as the descending path of bifurcating for locating M^{obs}, we denote the left and right branches at k-th bifurcation as $L_{d_k^o}$ and $R_{d_k^o}$ with $k = 1, ..., K_o$. The sizes of the two branches are denoted as $|L_{d_k^o}|$ and $|R_{d_k^o}|$. Then the size of the branch containing M^{obs} at k-th bifurcation is calculated as

$$|L_{d_k^o}|^{(1-d_k^o)}|R_{d_k^o}|^{(d_k^o)}. \tag{2.1}$$

Then there is an odds of correctly guessing which branch contains M^{obs} is calculated as:

$$PO[d_k^o|M^{obs}] = \frac{|L_{d_k^o}|^{(1-d_k^o)}|R_{d_k^o}|^{(d_k^o)}}{|L_{d_k^o}|^{(d_k^o)}|R_{d_k^o}|^{(1-d_k^o)}}. \tag{2.2}$$

We then compute the overall odds of guessing correctly on which branch M^{obs} belongs along the entire coding sequence $< d_1^o, d_2^o, ...d_{K_o}^o >$ as

$$PO(M^{obs}) = \prod_{k=1}^{K_o} PO[d_k^o|M^{obs}]. \tag{2.3}$$

An example is illustrated in Fig 2.12.

FIGURE 2.12. An example of the product of odds and p-value.

Likewise we compute an ensemble of odds $\{PO(M_b)\}_{b=1}^{B}$. Then we compute the p-value of observing an odds like $PO(M^{obs})$ as the proportion of $PO(M_b)$ being less than $PO(M^{obs})$:

$$p(M^{obs}) = \frac{\sum_{b=1}^{B} 1_{[PO(M_b)<PO(M^{obs})]}}{B}. \tag{2.4}$$

Upon the HC tree shown in Fig 2.11B, we have $PO(M^{obs}) = 0.023$ and p-value is $p(M^{obs}) = 0.002$. Hence, it turns out that M^{obs} is significantly extreme in the HC tree. Based on the visible patterns observed in the heatmap, we can conclude that the 25 [R1] (Highly-dense) rectangles are not uniformly distributed.

2.4.4. Enlarge the spatial scale: In the previous section, we have built a MST for the Highly-dense rectangles' 2D-distribution, as the 3rd of the flow chart shown in Fig 2.9C. We further work on the 2D distribution of the 4th one including those Dense rectangles, as shown in Fig 2.13A. By combining the [R1] and [R2] scales of rectangles, we expect to see the distribution of purple-dots with an expanded perspective. Its MST is computed and reported in Fig 2.13B. Likewise, we perform the two versions of testing on spatial uniformness and report the results in Fig 2.14, respectively. Though having less significance in terms of p-values, the results via ROC curve analysis on the left panel and Machine Learning method on the right panel all still indicate that the rectangles' 2D distribution is not fully in accord with spatial uniformness.

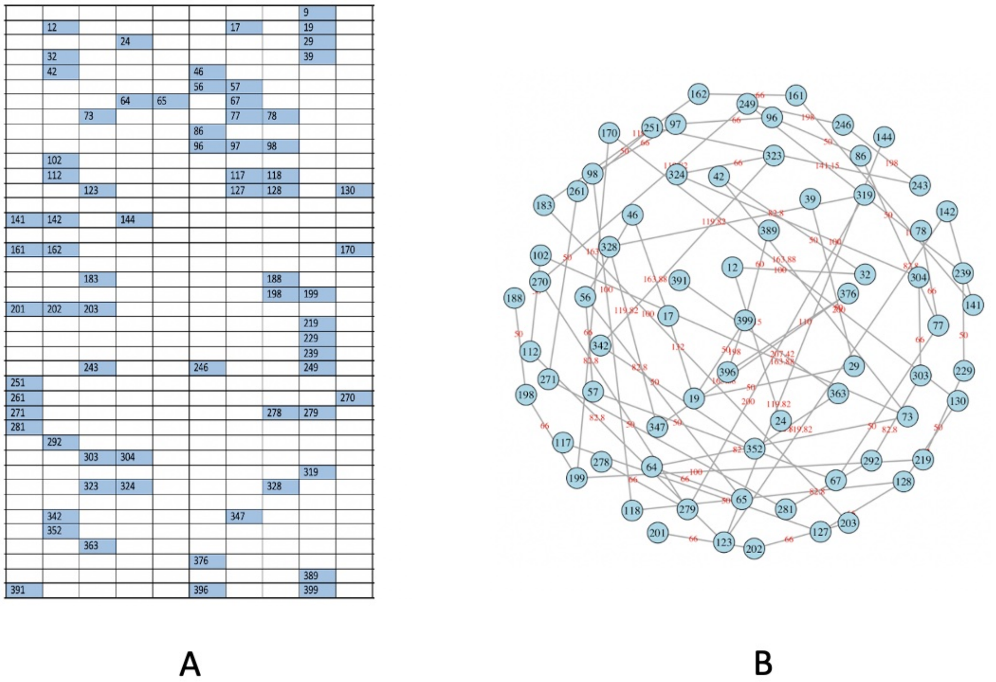


FIGURE 2.13. Enlarge the spatial scale.
A: focusing on [R1-R2] rectangles; B: the corresponding MST.

Nonetheless, this trend of getting less and less significant against spatial uniformness is expected when we further expand by including rectangles of [R3] Sparse scale. This trend tells us that the spraying mechanism needs further fine-tuning in order to achieve spatial uniformness. Especially, large purple nodes in [R1] Highly-dense scale should be significantly reduced.

2.4.5. Simulation study. In addition to the application to the real data, we conduct experimental simulations to validate our proposed method for 2D spatial uniformness testing via rectangle (or sub-region) neighborhood given known spatial scale densities. Two situations are considered in Fig 2.15. We divide the area into 400 rectangles. In simulation 1 (Fig 2.15A), all rectangles at spatial scale [R1-R4] are randomly uniformly distributed. In simulation 2 (Fig 2.15B), we select rectangles at [R1] around the center and bottom. Rectangles connected to those at [R1] are chosen as [R2] with a probability of 0.4. Similarly, rectangles connected to [R2] are selected as [R3] with a probability of 0.5. We sample [R4] rectangles from those connected to [R3] with a probability of

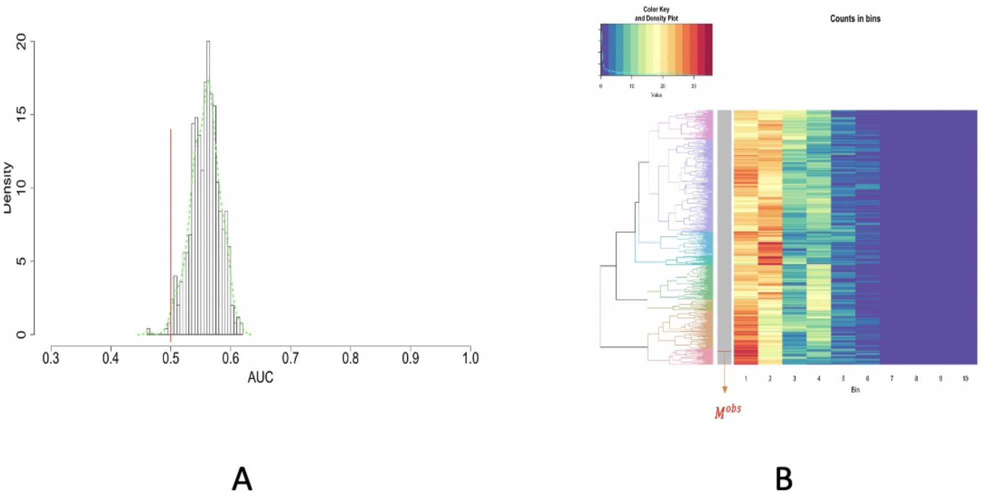


FIGURE 2.14. ROC curve and Unsupervised machine learning approach ontesting the spatial uniformness of [R1-R2] rectangles.
A: ROC curve analysis results; B: HC algorithm with hearmap, product of odds ($PO = 1.108$) and p-value $= 0.056$.

0.6. For each case, we apply our algorithm to three scales: Highly-dense ([R1]), Highly-dense with Dense ([R1-R2]), and Highly-dense with Dense and Sparse ([R1-R3]). The results are reported in Table 2.1 and consistent with our simulated data. The p-values and PO of three spatial scales in simulation 1 are all relatively large, indicating the uniformness of purple distribution. In simulation 2, a non-uniformness distribution is suggested according to small p-values.

TABLE 2.1. Computed results for two simulations. p-values with PO in parentheses are reported.

	[R1]	[R1-R2]	[R1-R3]
Simulation 1	0.558 (20397.38)	0.418 (64.48407)	0.934 (222504.8)
Simulation 2	0.008 (0.03275434)	<0.0001 (0.002)	<0.0001 (0.002)

2.5. Analysis of 4 other images

We apply computational algorithms developed and illustrated through image no.1 to the rest of the four images. Heterogeneous shading conditions can be evidently seen across these four images. Overall our exhaustive color identifications are satisfactory and the testings of spatial uniformness are indeed much more effective than the one based on ROC analysis.

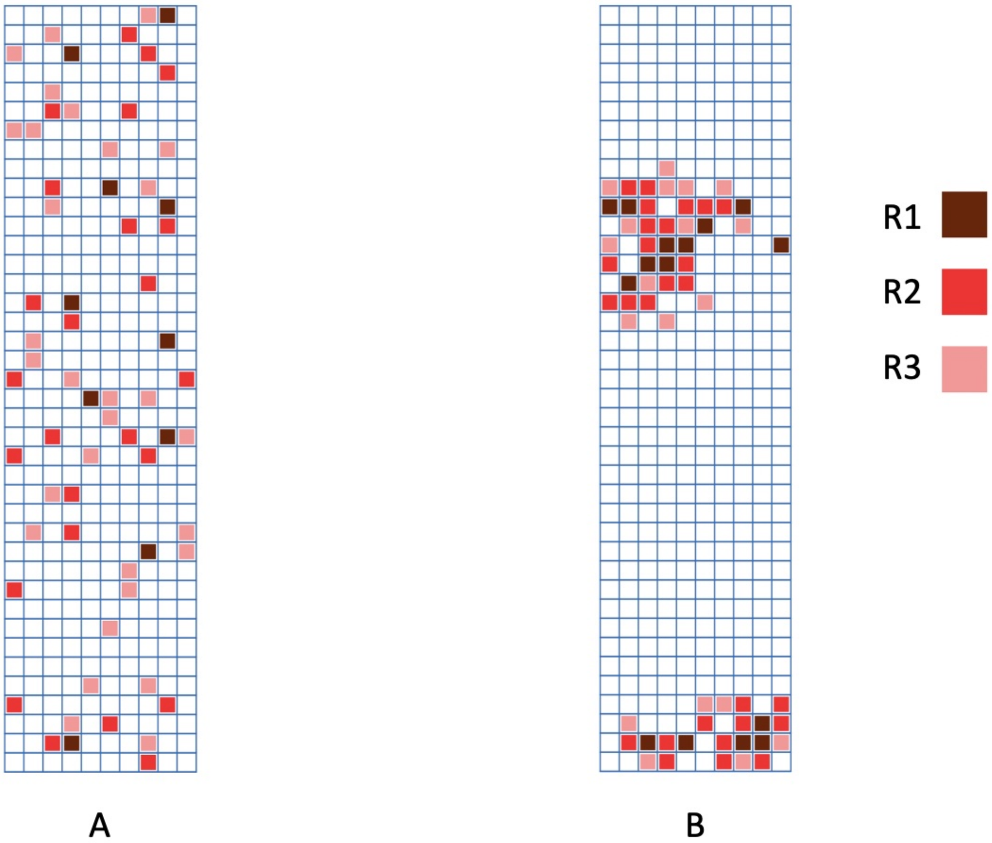


FIGURE 2.15. Two simulations. A: simulation 1 - purple rectangles are uniformly distributed; B: simulation 2 - purple rectangles are clustered in two sub-regions.

2.5.1. Image no.2. Image no.2 consists of two "pages" of test papers, as shown in Fig 2.16A. The upper part of these two coupled test papers is under shading. Consequently, the color-dot identification based only on RGB has missed quite a few small purple dots, as shown in Fig 2.16B. Many of these small dots were also not picked up via HSV data format based on $1 \times 1 \times 1$ fine-scale cubes.

Based on the observation as well as p-values indicated, the purple dots are not distributed uniformly and improvement in spraying via drones is needed. We separately report results of spatial uniformness on the two test papers by focusing only dense squares as shown in Fig 2.16C. Two separate results are reported: Fig 2.17 for the Left and Fig 2.18 for the Right, respectively. Based on both figures, we see that there exists a small discrepancy in p-values between the result based on ROC in Fig 2.17A against results based on HC-tree and heatmap in Fig 2.17B, and result

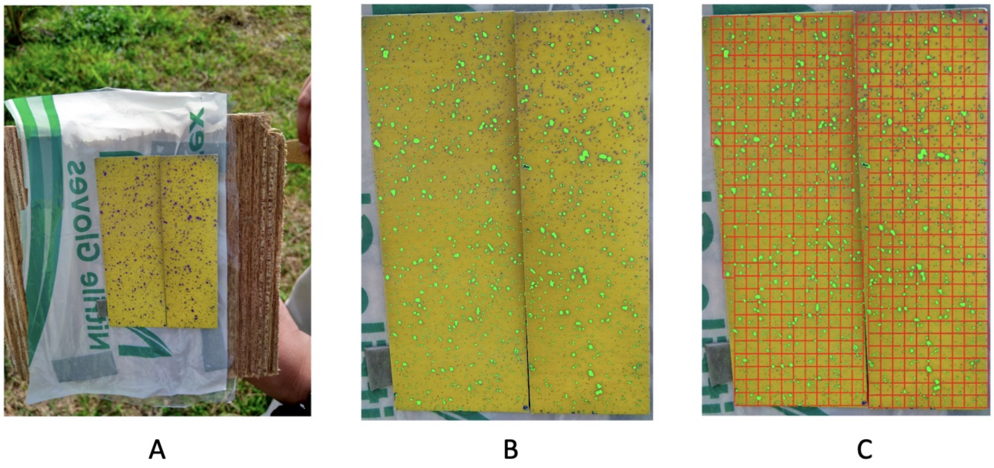


FIGURE 2.16. Image no.2:
A: two test papers in the original data; B: recovering by RGB file; C: dividing the paper into rectangles for 2D spatial uniformness testing.

based on ROC in Fig 2.18A against results based on HC-tree and heatmap in Fig 2.18B. However, such small discrepancies don't seem to cause incoherent conclusions.

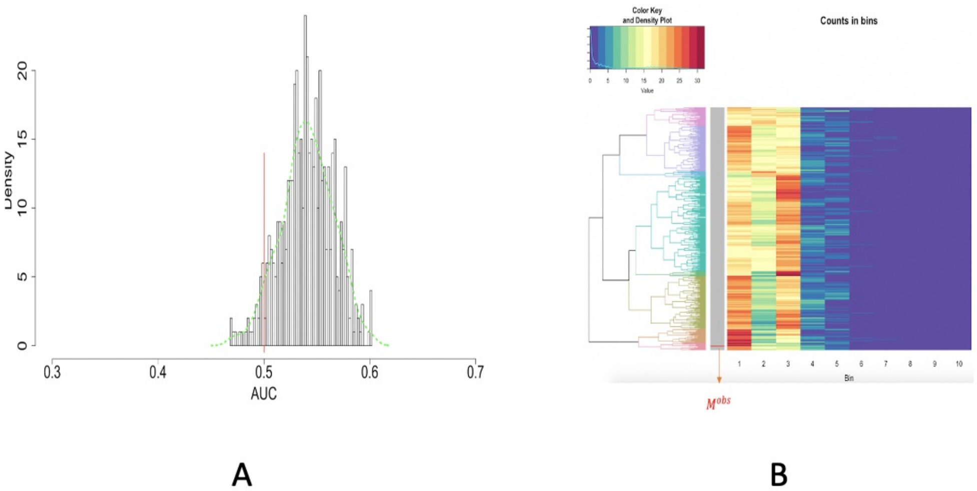


FIGURE 2.17. Image no.2: Spatial uniformness testing on the LEFT test paper: studying [R1] rectangles; A: ROC curve anlysis results; B: HC algorithm with heatmap, product of odds ($PO = 0.016$) and p-value $p(M^{obs}) = 0$.

2.5.2. Image no.3. A conclusion of uniform distribution of purple dots can be confirmed, according to the p-value. The test paper in image no.3 seems curved a bit, as shown in Fig 2.19A. This curved shape likely created shads around the upper left and lower right corners of the test

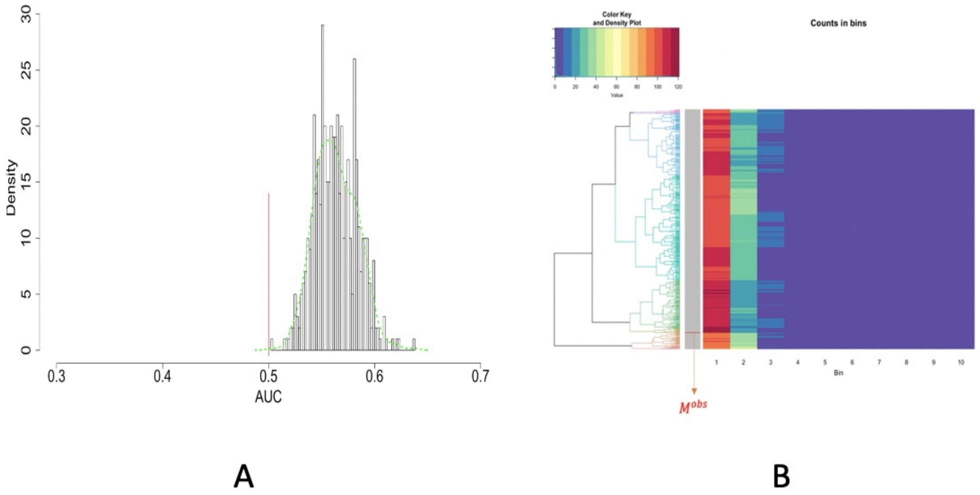


Figure 2.18. Image no.2: Spatial uniformness testing on the RIGHT test paper: studying [R1] rectangles; A: ROC curve anlysis results; B: HC algorithm with heatmap, product of odds ($PO = 0.047$) and p-value $p(M^{obs}) = 0.004$.

paper. The coupled results from RGB and HSV seem to achieve a big degree of exhaustive identification except dots locating the two corners, as shown in Fig 2.19B. For the spatial uniformness test, the result based on ROC analysis, as shown in Fig 2.20A, seems to point to a direction being slightly different from the one based on HC-tree and heatmap, as shown in Fig 2.20B. Given the observed row being well mixed with simulated one within a big branch, we are more confident on the p-value result based on HC-tree and heatmap Fig 2.20B. According to these results, we can conclude that the drone is satisfying.

2.5.3. Image no.4. Image no.4 has obvious shades at the upper and lower boundaries of the test paper, as shown in Fig 2.21A. We report the color-dot identification result based on HSV cubes of $n = 10$ scale, as shown in Fig 2.21B. For spatial uniformness testing, a big gap is seen between the result based on ROC analysis and the one based on HC-tree and Heatmap, as shown in Fig 2.22A and 2.22B, respectively. Again the latter result seems more reliable. Consequently, we are satisfied with even spread of purple dots based on a relatively large p-value.

2.5.4. Image no.5. Image no.5 looks like it is being twisted a bit, particularly at the lower right corner, and the shades are visible all over, as shown in Fig 2.23A. The final color-dot identification is based on the coupled results of RGB and HSV, as shown in Fig 2.23B. The discrepancy

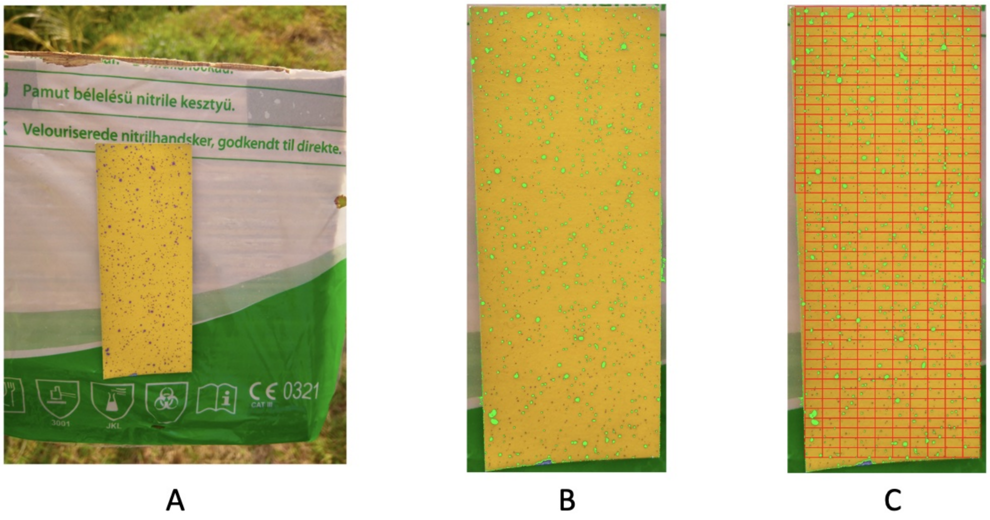

Pamut bélelésü nitrile kesztyü.
Velouriserede nitrilhandsker, godkendt til direkte.
CE 0321
A
B
C

FIGURE 2.19. Image no.3: A one test paper in the original data; B recovering by 6D [RGB+HSV] file ($n = 10$); (c) dividing the paper into 479 rectangles for 2D spatial uniformness testing.

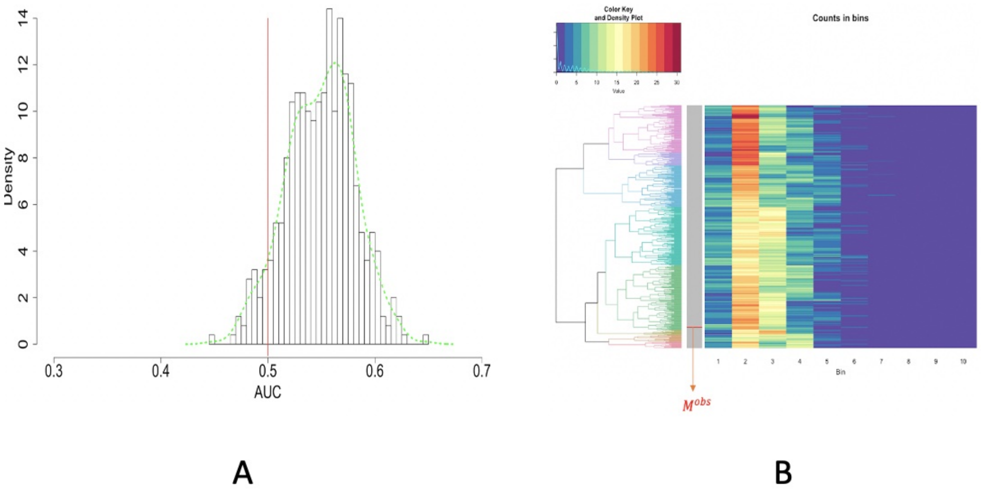


FIGURE 2.20. Image no.3: Spatial uniformness testing based on [R1] rectangles; A: ROC curve anlysis results; B: HC algorithm with heatmap, product of odds ($PO = 1164.056$) and p-value $p(M^{obs}) = 0.686$.

between the two results of spatial uniformness testing is especially wide. But, based on the small size of the branch containing the observed row vector, we have more confidence in the one based on HC-tree and heatmap, as shown in Fig 2.24B, over the ROC one, as shown in Fig 2.24A. The distribution of purple dots is marginally uniform because of the relatively small p-value, which indicates a fair job done by the drone.

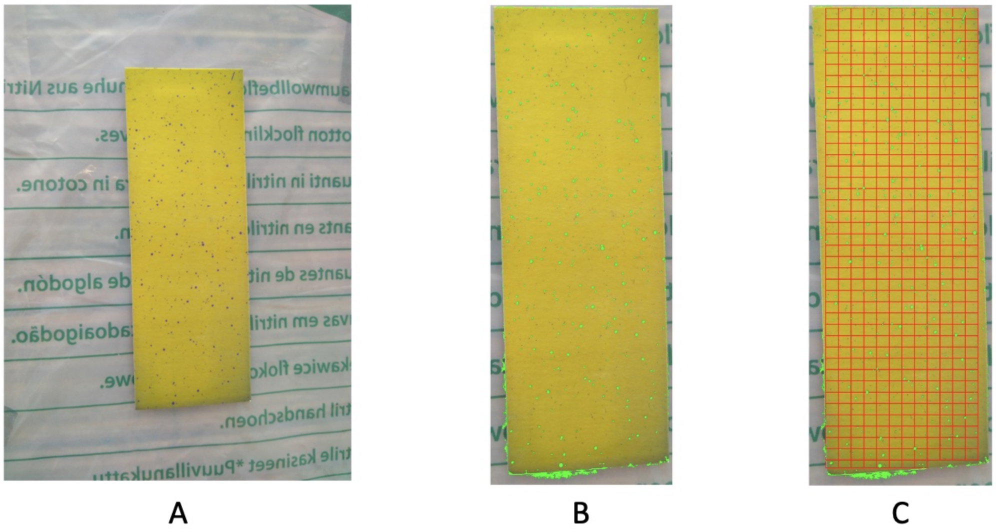


Figure 2.21. Image no.4: A: one test paper in the original data; B: recovering by HSV file ($n = 10$); C: dividing the paper into 488 rectangles for 2D spatial uniformness testing.

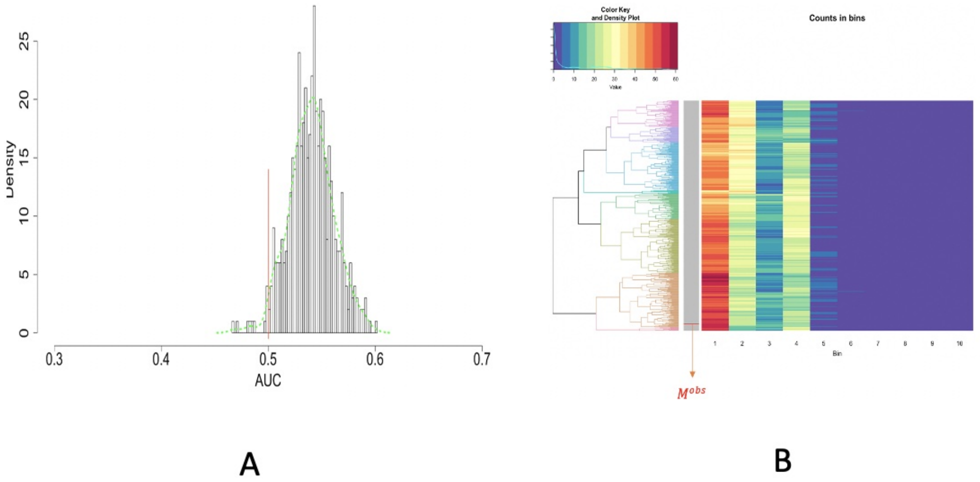


Figure 2.22. Image no.4: Spatial uniformness testing based on [R1] rectangles; A: ROC curve anlysis results; B: HC algorithm with heatmap, product of odds ($PO = 193.157$) and p-value $p(M^{obs}) = 0.642$.

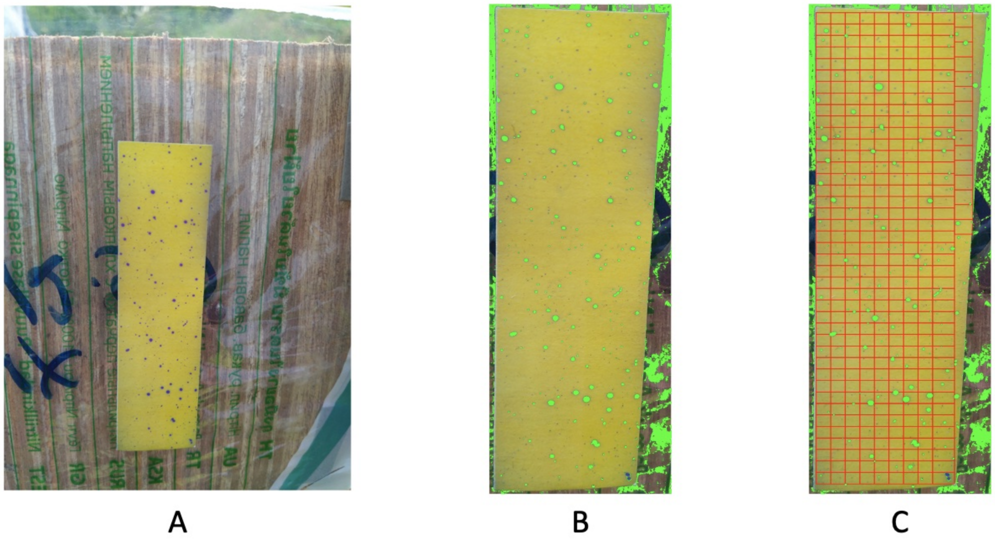


FIGURE 2.23. Image no.5: A: one test paper in the original data; B: recovering by 6D [RGB+HSV] file ($n = 10$); C: dividing the paper into 384 rectangles for 2D spatial uniformness testing.

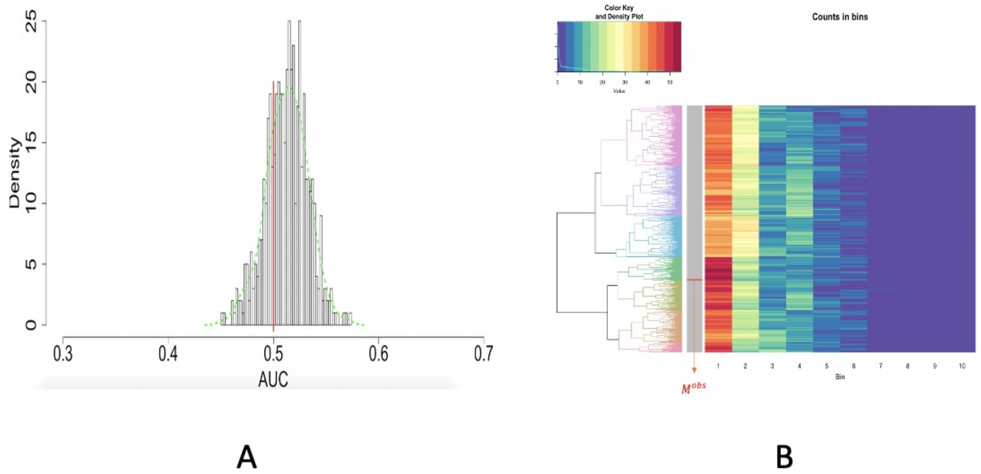


FIGURE 2.24. Image no.5: Spatial uniformness testing based on [R1] rectangles; A: ROC curve anlysis results; B: HC algorithm with heatmap, product of odds ($PO = 0.334$) and p-value $p(M^{obs}) = 0.056$

2.6. Conclusions

The color-identifications and testing 2D spatial uniformness via MST for the five images are rather satisfactory. Basically, these results collectively strongly indicate that our data-driven computational approach for color-identifications is rather effective, and testing methodology for 2D spatial uniformness is novel and practical.

The underlying reason for the effectiveness of our color-identification approach is the low color-complexity. This interesting fact is that this simple concept is not well known in the literature. In fact, our current color research has shown us that low color-complexity is seen in natural images as well as in images of famous paintings. That is, our data-driven color-identification is applicable in a wide range of color images.

Admittedly, employing stereo cameras with more single imaging angles is able to provide a better observation, even help eliminate the shading effect. Such Data can reveal the truly 3D-homogeneity of 3D-spatial distributions. Analysis of such 3D distribution data can resolve many realistic and essential issues facing precision agriculture, that is unlikely could be solved by combinations of RGB and HSV data of 2D images. However, provided with the limited 2D image-format in this paper, we try our best to extract as much information as we can from observed data sets. Given that RGB and HSV are two distinct aspects of the same images, they provide slightly distinct pattern information. Therefore, it is natural to combine both aspects of pattern information in hope of improving on results relying on solo representation.

The MST structure and its distance distribution are new and essential summarizing pattern information of spatial data. They are shown to be good approaches to illustrating and characterizing useful data structures in Data Science. The novelty of evaluating p-value via products of odds-ratios based on a tree structure, which is complex, can critically expand the applications of unsupervised machine learning methodologies to wider ranges of scientific fields, including the medical one.

Our way of dealing with shading in images is not sophisticated. We adopt the fact that RGB and HSV data formats could be differentially affected by shading. So, we propose to combine results from both data formats under distinct scales. From the results reported in Section 5, we see that it works with different degrees of success. It might be possible to develop systematic approaches to remove, or at least lessen the shading effects. This is one of our undertaking research directions right now.

CHAPTER 3

The geometry of colors in van Gogh's Sunflowers

3.1. Introduction

Vincent van Gogh's series of paintings of *Sunflowers* are among his most famous creations. There are two such series, the first one executed in Paris in 1887 that depicts sunflowers lying on the ground, while the second series was made later in Arles and represents sunflowers in a vase. Van Gogh made four versions of the Arles series in 1888, with different flower arrangements and different backgrounds: turquoise for the first version, F453, royal-blue for the second version, F459, blue-green for the third version, F456, and yellow for the fourth version, F454. Note that we use here the recognized classification of van Gogh's paintings, where F stands for the De La Faille catalogue of these paintings. A year later van Gogh made one repeat of the third version (F455), and two repeats of the 4th version (F458 and F457). The authenticity of the second repetition of the 4th version, F457, has been questioned, although experts now believe it is authentic [**103**]. These three repetitions and their two originals are usually referred to as the *Sunflowers in a vase*, where the number of sunflowers vary even in van Gogh's own recollection as described in his correspondence [**7**]. They are all currently housed in different museums: the third version, F456, is in the Neue Pinakothek museum in Munich, Germany, and its repetition, F455, is in the Philadelphia Museum of Art, United States; the fourth version, F454, is at the National Gallery in London, England, while its repetitions, F457 is at the Seiji Togo memorial Sompo Japan Museum of Art, Tokyo, Japan, and F458 is at the van Gogh Museum Amsterdam. Those paintings are rarely reunited, with possibly the latest occurrence being an exhibit at the National Gallery in London in 2014 were the original and one repetition of the 4th version were shown side to side. Interestingly, however, it is possible to see them virtually in the same room in a virtual 360° exhibition commented by Vincent van Gogh's grandson (`https://www.facebook.com/VanGoghMuseum/videos/10159187334010597/`). In this paper we focus on the fourth version, F454, and its two repetitions, F457 and F458, in which the vase with sunflowers is portrayed against a yellow background. This version is considered as a study of variations on the theme of yellow, with the aim of achieving a light-on-light effect [**102**, **103**]. Pictures of those three paintings are shown in Fig 3.1.

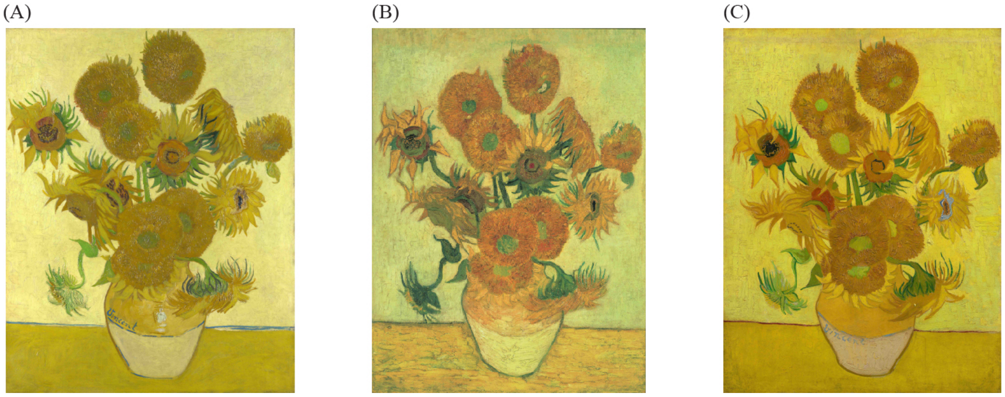


FIGURE 3.1. Pictures of the three paintings of *Sunflowers in a vase with a yellow background* from Vincent van Gogh: (A) the fourth version, F454, painted in 1888 in Arles and currently at the National Gallery in London, and its two repetitions, also painted in Arles in 1889, F457, currently in the Seiji Tojo memorial Sampo Museum in Tokyo (B), and F458, at the van Gogh museum in Amsterdam (C)

Van Gogh's *Sunflowers* appeal to many. The fascination exerted by van Gogh's stay in the south of France in 1888-1889, his friendship and fallout with Paul Gauguin, his mental health, and the seven sunflowers canvases he painted during that time remain a tremendous source of inspiration in many forms of popular entertainment as well as for artists, museum curators, and scientists in general. Biologists have studied how bees interact with those paintings [**16**], the genetic fabric of the sunflowers [**13**], doctors have attempted to associate his medical conditions and related treatments to his perception of colors [**6**, **11**, **41**]. Of direct relevance to the paintings themselves, scientists and museum curators have used a broad array of traditional and state-of-the-art techniques to look at and below the paint surface itself [**45**, **52**, **80**, **81**, **82**, **83**, **84**, **100**, **104**]. Of special interest, we advise the reader to consider the recent book, *Van Gogh's Sunflowers Illuminated* published by the Amsterdam van Gogh museum that summarizes the results of the research undertaken by an international team of scientists, curators, and art historians aimed at comparing the F454 *Sunflower* from the London National gallery with one of its repetitions, F458 from the van Gogh museum [**46**]. In this paper, we approach the same problem of comparing F454 with its repetitions F457 and F458 from a very different perspective. Instead of analyzing the canvases and the paint surfaces directly, we study high resolution images of the paintings using a data science approach. The availability of large collection of digital images of paintings has opened the door to the use of state-of-the-art supervised machine learning techniques in art, for understanding the artist style [**101**], for classifying paintings [**94**], to detect forgeries [**90**], and possibly for even creating art [**18**]. Our approach to analyzing one of the *Sunflowers* painting and two of its repeats differs as it is unsupervised. We consider a painting based on a high-resolution image of it. This image is a collection of pixels, with each pixel characterized by its location and color. The color is quantified based on the RGB color model. This is an additive color model in which Red, Green, and Blue lights are added together to reproduce any colors. In a digital image, the amount of each R, G, and B color is discrete, usually an integer in the range $[0, 255]$. As such, a pixel in an image belongs to a discrete color space of size $[0, 255]^3$. Our analysis of F454, F457, and F458 amounts to comparing their representations in this discrete color space. We are particularly interested in the sunflowers themselves, with their yellow crowns and stems as they are at the heart of the three paintings.

We acknowledge that our analyses may be somewhat subjective. We only use indirect representations of the three paintings by van Gogh, namely digital images of those paintings. While those images are high resolution, it is not impossible that there are some color distortions, although those are expected to be small with modern digital cameras (this will be addressed briefly in the following). More importantly, the paintings themselves are more than 130 years old and the modern images we have of those paintings certainly differ from their original renderings. It is interesting that van Gogh himself was well aware that paintings age, as he wrote to his brother on April 30th 1889 that “paintings fade like flowers” [**7**]. Many studies have focused on the chemistry of aging of van Gogh’s paintings [**45**, **80**, **81**, **82**, **83**, **84**, **100**, **104**], as well as attempts to predict the evolution of the appearance of those paintings [**39**, **44**, **62**]. With this as a background, there are two main directions that we have followed in our analyses of the digital images of F454, F457, and F458:

i) *Identify markers of aging from the color distributions in the digital images.* We were able to identify shifts in the red, blue, and green components of the images that are most likely a result of aging. The most significant shift is observed for the blue component of yellow hues, leading to a browning and fading of those colors. Using this shift, we propose pseudo-color reconstruction schemes that enable us to generate model images of the original paintings. We note that the same procedure can be used to extrapolate to models of the paintings in the future.

ii) *Compare and contrast the use of colors in the original version, F454, and the two repetitions, F457 and F458, of the* Sunflowers in a vase with a yellow background. In particular, we identify differences between the repetitions that most likely reflect different intents by van Gogh.

We organize the paper as follows. Major findings are shown in Section 3.2. In particular, we provide an in depth analysis of the RGB color spaces of the digital images corresponding to the three paintings, F454, F457, and F458. For clarity reason, we will refer to F454 as being the "London version", F457 as the "Tokyo version", and F458 as the "Amsterdam version". By comparison with recent images of actual sunflowers, we identify significant shifts in the blue components, B, of the images of those paintings and use those shifts to derive pseudo-color reconstruction (PCR) schemes to attempt to restore the original images. We also propose a color transfer scheme that enables us to compare the three paintings. In the conclusion section 3.2 we briefly discuss the benefits of our approach as well as future directions of research. Finally, in Section 3.4 we provide technical details on the images we have used, on the statistical methods that we implemented, as well as on the algorithms we have implemented.

3.2. Results and Discussion

3.2.1. Basic statistics on the color content of the three digital images of the *Sunflowers* paintings. A digital image is composed of basic image elements, i.e. pixels. Each pixel is characterized by its position in the image and its color. In the RGB model, a color is described by 3 components, also called coordinates, representing the amounts r, g, and b, of red, green, and blue that need to be combined to generate that specific color. Each coordinate is an integer value taken in the range $[0, 255]$. There are therefore 256^3 (approx. 17 millions) possible colors.

We have used high resolution images for all three Sunflower paintings, namely the London version, the Tokyo version, and the Amsterdam version. We are aware that there might be some color distortion in those images, associated with the fact that it is likely that those images were taken with different cameras. In addition, we cannot exclude the effect of the resolution. While it is difficult to assess the importance of the cameras, we can at least assess the importance of the resolution. We generated the distributions of the values for the red, green, and blue coordinates over the three high resolution digital images of the three paintings, as well as for corresponding low resolution images (see Section 3.4 for details on how to generate those distributions). Results are shown in Figure 3.2. The low resolution images were taken from the museum websites directly. For the London version, the high resolution image has 3349×4226 pixels, while its corresponding low resolution image has 629×800 pixels. Similarly, the high resolution image of the Tokyo version has 3626×4829 pixels, while its corresponding low resolution image has 450×602 pixels, and the high resolution image of the Amsterdam version has 3324×4226 pixels, while its corresponding low resolution image has 609×800 pixels. It is interesting that despite those significant differences in resolution, the distributions of red, green, and blue are very similar (although admittedly not identical) for the high of low resolution images, for all three paintings, while different when comparing the paintings themselves. This gives us some confidence that we can perform such comparisons. In the remainder of the paper, we will use only the high resolution images.

As mentioned above, the distributions of red, green, and blue differ between the three paintings (see Figure 3.2). While the distributions of red and green values are relatively similar over the three images, with values that concentrate in the range $[100, 255]$, the distributions of values of the blue coordinate differ significantly. In the Tokyo version there are basically no pixels whose blue intensity is below 60 or otherwise stated nearly all its pixels have some blue component. In contrast, in the Amsterdam version there are more than 5 % of the pixels that have a blue coordinate close to zero, namely no blue component. Clearly, something is different with the contributions of blue in the paintings. This will be discussed in detail below.

Interestingly, in the high resolution digital images of the three paintings, we observe only 2 %, 5 %, and 3 % of those 256^3 colors, respectively, indicative of relatively low diversities of colors in the three paintings. While this may not be surprising as all three paintings are predominantly yellow (see Figure 3.1), it is noteworthy that the color diversity differ significantly between the painting. Indeed, the Tokyo version and the Amsterdam version are defined as repetitions of the London version and therefore we could expect more similarities.

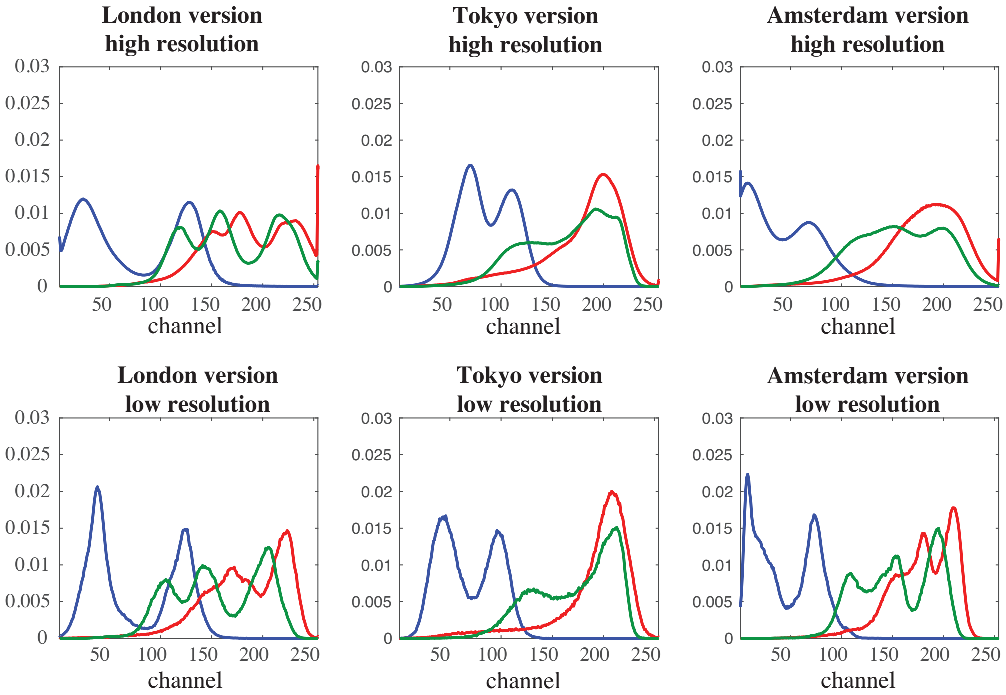


FIGURE 3.2. Distributions of the values of the color coordinates for Red (in red), Green (in green) and Blue (in blue) for all pixels of the high resolution images (top row) and of low resolution images of the three *Sunflowers* paintings, the London version, the Tokyo version, and the Amsterdam version. See material and methods for the provenance of the high resolution images. The low resolution images were obtained directly from the corresponding museum websites.

We looked also at the association between pairs of fundamental colors in all three paintings, by computing their relative conditional entropy HR (see Section 3.4 for details). The conditional entropy HR measures the amount of information shared between two colors. Our definition of HR places it in the interval $[0, 1]$; it is equal to zero if the two colors are fully determined by each other, while conversely it is equal to 1 if the two colors are completely independent. In Figure 3.3 we display the values for the nine pairs of colors in the form of tables for the three images corresponding to the London, Tokyo, and Amsterdam versions. As expected, a color compared to itself leads to a conditional entropy of zero. In contrast, the HR values for pairs of different colors reveal very low dependence between them. There are, however, noticeable differences between the three images.

3.2.2. Color distributions in the background, the sunflower crowns, and the sunflower stems of the *Sunflowers in a vase with a yellow background*. Visual inspections

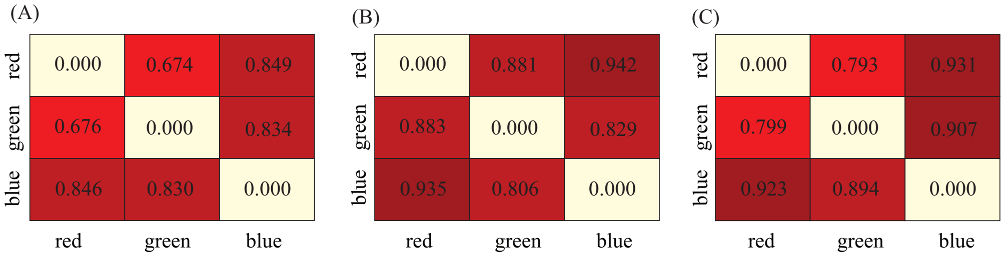


FIGURE 3.3. Relative conditional entropy values between the red, green, and blue coordinates of all pixels of (A) the London version, the Tokyo version, and (C) the Amsterdam version of the *Sunflowers in a vase with yellow background*; values range from 0 to 1 where 0 indicates dependence and 1 independence between the two coordinates.

show differences between the London version and the Tokyo and Amsterdam versions, respectively, especially when it comes to their colors. We see that the green color of the stems is similar to the expected green for stems of real sunflowers. On the other hand, the yellow flowers are darker than the expected vibrant yellow of the petals of a real sunflower, especially in the London version (see Figure 3.1). There are many possible reasons for this darker color of the flowers. First, it could have been Van Gogh's intent. It is known that Van Gogh painted his series of paintings on sunflowers using real sunflowers are models. He painted the London version in late August 1888, when sunflowers usually start to fade [**8**]. Second, van Gogh used the yellow chrome pigment to color the sunflowers and it is known that this pigment darkened as it ages [**9**, **82**, **84**]. Finally, we analyze the colors of the paintings using digital images of those paintings. The images are influenced by the lighting that was used at the time that they were taken, which may influence our perception of their colors (see discussion above on the importance of resolution, as well as reference [**62**]). It is therefore difficult to isolate a specific origin, especially as those mentioned above are ultimately non-exclusive. We propose a data-driven approach based on the analysis of the color space of the three paintings to provide some quantitative elements that describe the differences between the three paintings, and their differences with real-life sunflowers.

The *Sunflowers in a vase with a yellow background* paintings include 15 sunflowers with their stem in a vase standing on a table, with a yellowish background. For each painting, we identify three regions of interest, or ROI. A ROI is defined according to its position in the image as well as from its color consistency. The first two ROIs relate to the flower crowns of the sunflowers (mainly yellow, yf-ROI) and the stems of the sunflowers (green, g-ROI), while the third ROI refers to the background (bg-ROI), which includes the table, vase, and the region behind the sunflowers. In parallel, we analyzed two recents photographs of sunflowers (see Figure 3.4), from which we extracted the corresponding flower ROIs and stem ROIs. Extractions of the ROI was performed as described in the Material and Methods section. Each ROI is defined as a set of pixels, characterized by their locations and colors, with the latter given in the RGB space with three discrete coordinates with values in the range $[0, 255]$.

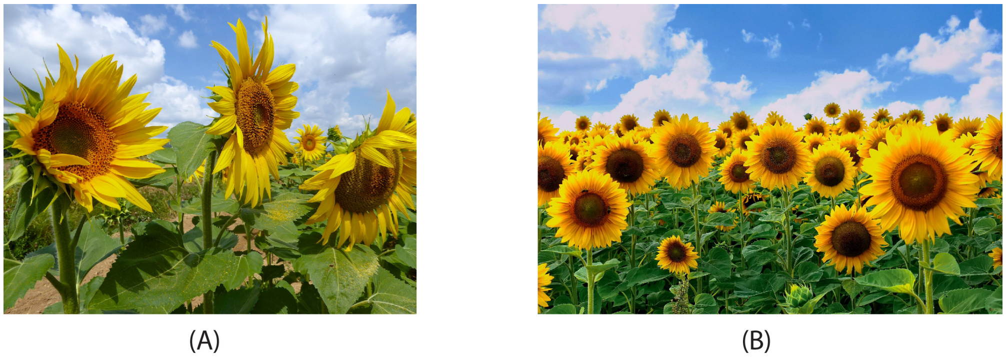


FIGURE 3.4. Two images of natural live sunflowers (see text for details): (A) SF1; (B) SF2.

The yellow flowers of the *Sunflowers in a vase with a yellow background*. We compared first the ROIs associated with the yellow parts of the sunflowers of the three paintings. Figure 3.5 shows the distribution of colors of the pixels associated with those ROIs in the red-green, red-blue, and green-blue planes. In the red-green plane, most pixels are found to follow the first diagonal, consistent with the general yellow color of those ROIs. Interestingly, the original, namely the London version, and the Amsterdam repetition are found to be very similar. Both differ from the Tokyo repetition. The differences with the Tokyo version are striking in the red-blue and blue-green planes, as blue appears with a wide range of values in the London and Amsterdam versions, from 0 to 120, while it is always present in the Tokyo version, appearing in a much smaller range centered around 60.

While there are differences between the three paintings in the diversity of color of their yellow sunflowers, there are even bigger differences when we compare those sunflowers with real sunflowers,

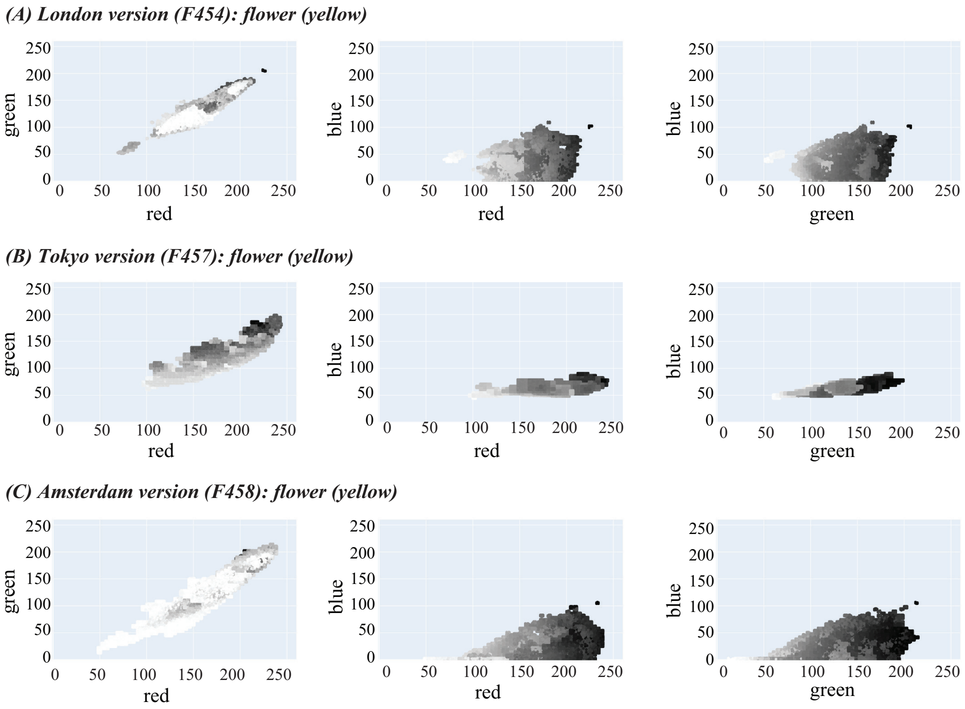


FIGURE 3.5. The RGB space occupied by the pixels of the yellow flower ROIs of the original *Sunflowers in a vase with yellow background*, the London version (top row), compared with the two duplicates the Tokyo version (middle row), and the Amsterdam version (bottom row). (A) red vs green; each dot represents one pixel; its color (grayscale) suggests the intensity of its blue coordinate, ranging from low (white) to high (black); (B) red vs blue; (C) green vs blue

as observed by comparing Figure 3.5 with Figure 3.6. Those differences can be summarized as follows:

i) The yellow of real sunflowers contain a minimal amount of blue, as illustrated in the blue-red and blue-green planes for the two photographs SF1 and SF2. In contrast, as indicated above, the yellow observed in the sunflowers of the paintings contains a significant amount of blue.

ii) The yellow of the flowers of real sunflowers contains a wide range of green covering nearly the whole spectrum of possible values from 0 to 250. In contrast, the green in the yellow of the flowers in the paintings is limited to the range $[50, 250]$ in all three paintings.

iii) Similarly, the yellow of the flowers of real sunflowers contains a wide range of red with values between 20 and 255. In contrast, the red in the yellow of the flowers in the paintings is limited to the range $[100, 250]$ in all three paintings.

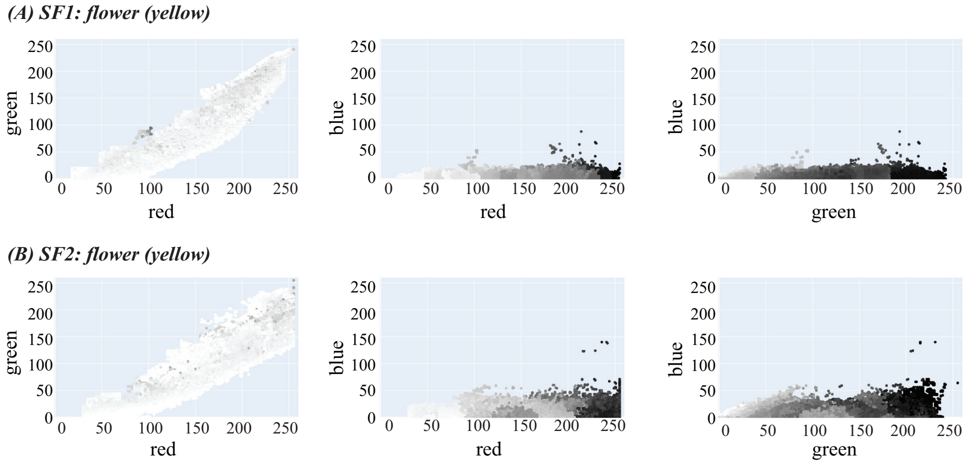


FIGURE 3.6. The RGB space occupied by the pixels of the yellow flower ROIs of the photographs SF1 (top row) and SF2 (bottom row) of real sunflowers. (A) red vs green; each dot represents one pixel; its color (grayscale) suggests the intensity of its blue coordinate, ranging from low (white) to high (black); (B) red vs blue; (C) green vs blue

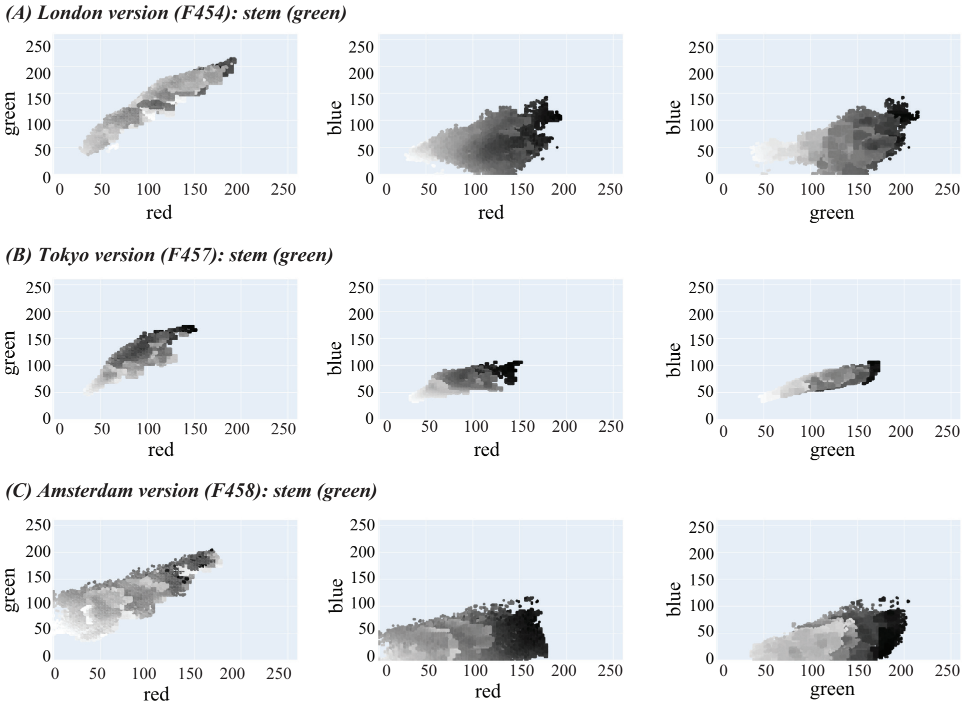
The stems of the *Sunflowers in a vase with a yellow background*. We repeated the analysis described above on the ROIs corresponding to the green stems of the sunflowers, both for the 3 paintings of sunflowers and for the two photographs of real sunflowers. Results are shown in Figure 3.7 and Figure 3.8, respectively. Comparisons of those two figures lead to the following observations:

i) Among the two copies, namely the Tokyo version and the Amsterdam version, the latter is the closest to the original (the London version). This result is the same as what was observed for the yellow parts of the petals and crowns of the sunflowers. Differences associated to the Tokyo version again come from a smaller range of blue coordinates for the pixels representing the stems.

ii) On average, the pixels in the green ROI of the Tokyo version contain more blue than the corresponding pixels in the London version and the Amsterdam version. This is visible in the paintings themselves, as the stems in the Tokyo version appear with darker green (a consequence of the addition of blue) than those of the London and Amstradam versions.

iii) There is a stronger linear relationship between red and green coordinates for the stems of the real sunflowers, compared to the stems in the three paintings.

iv) Besides the stronger relationship indicated above, the colors of the stems of the painted sunflowers qualitatively resemble the colors of the stems of real sunflowers, as captured by photographs.


(A) London version (F454): stem (green)

(B) Tokyo version (F457): stem (green)

(C) Amsterdam version (F458): stem (green)

FIGURE 3.7. The RGB space occupied by the pixels in the green stem ROIs of the original *Sunflowers in a vase with a yellow background*, the London version (top row), compared with the two duplicates, the Tokyo version (middle row), and the Amsterdam version (bottom row). (A) red vs green; each dot represents one pixel; its color (grayscale) suggests the intensity of its blue coordinate, ranging from low (white) to high (black); (B) red vs blue; (C) green vs blue

The backgrounds of the *Sunflowers in a vase with a yellow background*. Finally, we compared the backgrounds of the three paintings in Figure 3.9. The background ROIs are expected to be the most diverse as they include multiple parts of the paintings. In addition, visual inspections of those paintings clearly indicate significant differences (see Figure 3.1): the background of the

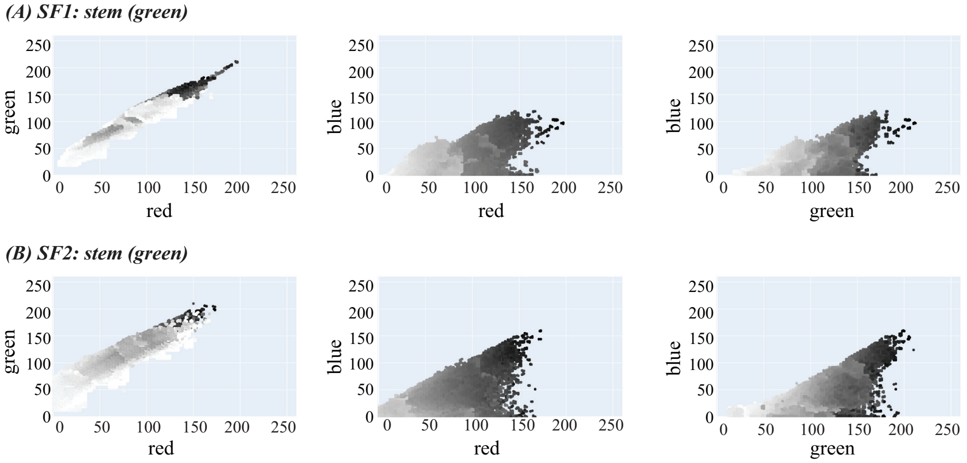


Figure 3.8. The RGB space occupied by the pixels of the green stem ROIs of the photographs SF1 (top row) and SF2 (bottom row) of real sunflowers. (A) red vs green; each dot represents one pixel; its color (grayscale) suggests the intensity of its blue coordinate, ranging from low (white) to high (black); (B) red vs blue; (C) green vs blue

London version is globally pale yellow, while those of the Tokyo and Amsterdam versions are more green-yellow and solid yellow, respectively. Those differences are reflected in the differences between the RGB space occupied by the pixels of the background, as illustrated in Figure 3.9. The background ROI within the London version shows a large number of pixels with large R, G, and B coordinates; those pixels will show as close to white. In contrast, a significant number of pixels in the background of the Tokyo version have large G coordinates, consistent with a green coloration. Finally, the background of the Amsterdam version includes many pixels with high R and G coordinates, and low B coordinate, consistent with a solid yellow.

Why those differences between the paintings, and between the painted *Sunflowers* and real sunflowers? As mentioned in the introduction of this section, there are possibly three main reasons for differences between the *Sunflowers* paintings themselves, and between the paintings and real sunflowers. Those reasons are associated to the painter's intent, to the aging of the paintings, and to artifacts associated with the digital images we consider. We exclude the latter as we believe that such artifacts are minor compared to elements of the two other reasons [**62**]. The analyses we have presented above provide elements that highlight the importance of aging.

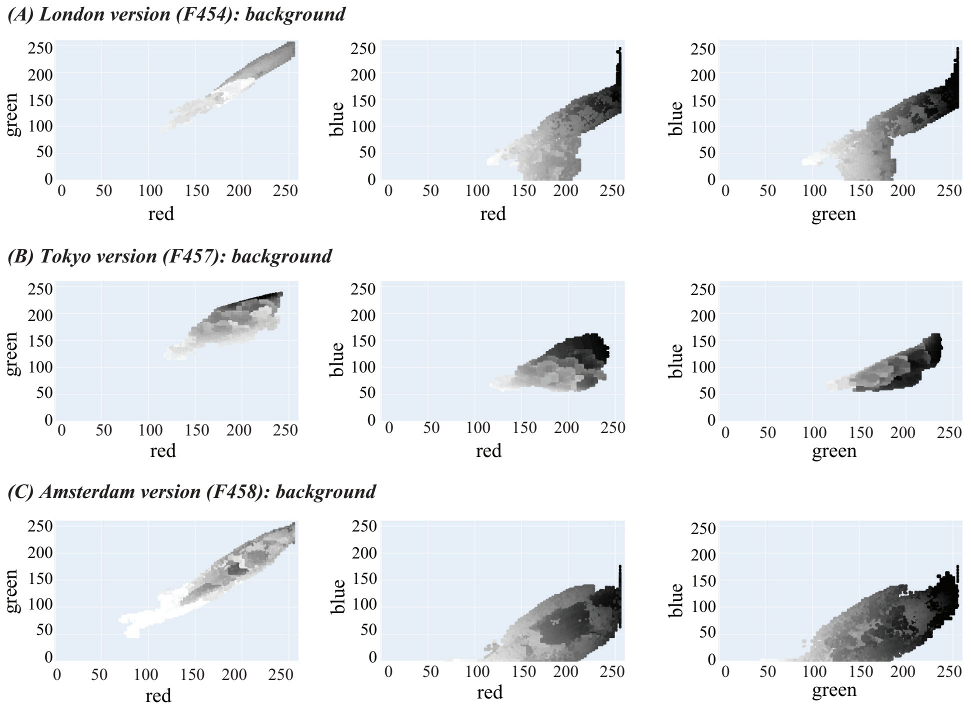


FIGURE 3.9. The RGB space occupied by the pixels in the background ROIs of the original *Sunflowers in a vase with a yellow background*, the London version (top row), compared with the two duplicates, the Tokyo version (middle row), and the Amsterdam version (bottom row). Note that the background includes the table, the vase, and the region behind the flowers. (A) red vs green; each dot represents one pixel; its color (grayscale) suggests the intensity of its blue coordinate, ranging from low (white) to high (black); (B) red vs blue; (C) green vs blue

The three Sunflower paintings currently in London, Tokyo, and Amsterdam are more than 130 year old and as such they have been aging. They have been subject to color degradation and deterioration as the color pigments are constantly exposed to light, humidity, pollution, and microbial contamination [**64**, **104**]. It is known that Van Gogh used commercial oil paints for his paintings. In a letter to Arnold Koning, dated January 1889 and believed to refer to the London version of the *Sunflowers* , van Gogh described them as being ”painted with the three chrome yellows, yellow ochre and Veronese green and nothing else” (letter 740 [**7**]). Recent X-ray fluorescence spectrometry analyses of the London version [**47**] and of the Tokyo version [**45**] revealed that indeed van Gogh was using chrome yellows to render yellow in those paintings; it is very likely that he was using the same pigments for the Amsterdam version. The aging of chrome yellow pigments used in paintings has been studied in details [**9**, **82**, **83**, **84**, **104**], including studies on van Gogh's *Sunflowers* [**80**]. Those studies have highlighted that the lightest hues in the chrome yellow family contain sulfate groups, which reduce the pigments' stability under light: bright yellow on canvases then turns to brownish green. Our comparisons of the yellow observed for real sunflowers and the yellow visible in the sunflowers in the three paintings are consistent with those observations. In particular, we observe an increase in the amount of blue for the yellow pixels associated with the flowers in those paintings, leading to a less vibrant yellow that may even look brown. In addition, we observe a shift to increased amount of green for those pixels (as none of them have green coordinates below 50), while the distribution of green in the yellow of real sunflowers covers the whole spectrum. This increase in blue and green is consistent with the yellow colors appearing more brownish green.

While the differences in the yellow hues of the painted *Sunflowers* compared to the yellow hues in real sunflowers are explicit for all three paintings, highlighting aging of those paintings, we have also observed significant differences between the Tokyo version of the repetition and the two other paintings, the London version (the original), and the Amsterdam version (the second repetition). This is most likely the intent of van Gogh, as proposed by Bakker and Ripoelle [**8**]. Indeed, van Gogh painted the Tokyo version not based on real sunflowers but based on another work of art, the London version (van Gogh painted the Tokyo version in January when there were no sunflowers available). In a letter to his brother Theo (letter 736, [**7**]), van Gogh mentioned that this repetition was meant to be "equivalent and identical", although it was clear that this referred to the subject (the vase and sunflowers), and not to details and colors. Van Gogh pushed chromatic intensity even further, with the aim of achieving a radical light-on-light effect, such that the green stalks of the flowers contrast even more strikingly with the various yellows than in the original, the London version. Those differences remain despite the aging of the paintings. In contrast, our analyses show that the Amsterdam version appears closer to the original, the London version, than the other repetition, the Tokyo version.

3.2.3. Can we remediate aging of colors in paintings digitally? Paintings age and consequently do not look today as they were originally designed by the artists. This is not evident at the level of the colors that undergo fading and sometimes changes in hue, such as yellow turning into brown. While this aging is inevitable, there is great interest among curators to recreate the artists' original colors to enhance the experience of museum visitors. Paintings, however, cannot be physically restored to their original colors and only reconstructions offer the possibility of recreating their appearance as intended by the artist. Over the past few years, digital reconstructions of several paintings by van Gogh and other post-impressionist artists have been published, including investigation of van Gogh's series of The Bedroom [**33**], The Starry Night [**108**], Undergrowth with Two Figures [**32**], Irises [**12**], Roses [**12**], and Fields with Irises near Arles [**62**]. Some of those reconstructions rely on identifications of regions within the paintings that contain pigments that may have aged, using X-ray fluorescence, followed by digital reconstruction using software that can manipulate images (see for example the digital reconstruction of van Gogh's "Undergrowth with two figures" [**32**], or applications of software for optimizing the display of images on mobile devices [**63**]. The analyses of the RGB space occupied by the pixels of the paintings *Sunflowers in a vase with a yellow background* provided above suggest two other methods for digital reconstruction, namely color correction and color transfer, which we illustrate below.

Restoring colors in van Gogh's *Sunflowers* using color transfer. One approach to correcting aging in a painting is to transfer color from a recent representation of an object onto the region representing that object in the painting. For example, we can transfer the yellow color from live sunflowers observed in photographs to the regions representing sunflowers in one of the paintings. We illustrate this process in Figure 3.10 for the London version, namely the original *Sunflowers in a vase with a yellow background.* We start with the ROI corresponding to the yellow flowers of the London version as well as the corresponding ROI of the yellow flowers in the photograph SF1. We apply hierarchical clustering on the pixels of each of those ROIs, using the Euclidean distance between their RGB coordinates as a metric, and complete linkage (see Methods). The hierarchical clustering generates a tree; we represent the leaves of that tree with vertical bars whose colors are the colors of the corresponding pixels. Figure 3.10 A and Figure 3.10 B illustrates the corresponding complete color bars for the yellow flower ROI of the London version, and for the yellow flower ROI of SF1, respectively. In the color bar associated to the London version, we identify the cluster with the darkest hues of yellow and label the corresponding pixels as F454_A. The positions of those pixels within the London version are highlighted in red in Figure 3.10 C. Similarly, we isolate the region with the brightest hues of yellow in SF1 and label corresponding pixels as SF1_A. Finally we transfer the color associated with SF1_A onto F454_A using the transfer algorithm described in the Method section. Result of the transfer is shown in Figure 3.10 E, to be compared with the original, the London version, shown in Figure 3.10 D. As expected, the sunflowers in the modified London version are much brighter.

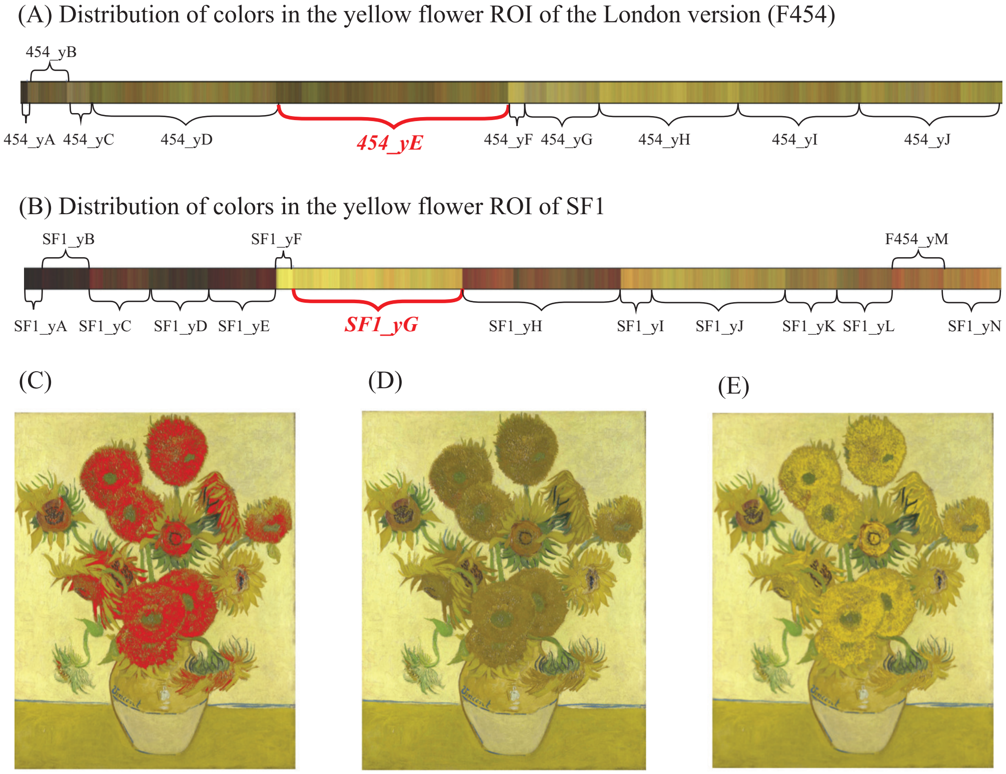


FIGURE 3.10. Color transfer from a real sunflower to the sunflowers in the London version (A) The color bar representing all hues of yellow in the ROI corresponding to the flowers in the London version. This color corresponds to the leaves of the HC tree computed from the colors of all pixels in this ROI. We select the region (cluster) 454_yE in this color bar which contains some of the darkest yellow pixels. The pixels associated to this region are referred to as F454_A. (B) The color bar representing all hues of yellow in the flower ROI of SF1. We identify the cluster SF1_yG in this color bar that contains some of the brightest yellow. The pixels associated to this region are referred to as SF1_A (C) Pixels in the London version that belong to F454_A are highlighted in red (D) London version: before color transfer (E) London version: after color transfer.

The color transfer strategy presented above can be expanded to play with color contrasts within the painting. We illustrate this concept by modifying the background of the flowers in the London version, as illustrated in Figure 3.11. van Gogh painted many variations of the sunflowers, with background varying from pale to deep blue and yellow as he explored chromatic effects in the juxtaposition of those backgrounds with the yellow flowers of the sunflowers [**8**]. We decided to modify at least part of this background to see how it visually impacts our perception of the paintings, using the London version as a support. We first identified pixels in the background of the London version that are located mostly behind the sunflowers and have light yellow hues. Those pixels are referred to as F454_B (see Figure 3.11 C). We then selected pixels in the real sunflowers depicted in the photograph SF1, which correspond to the dark yellow part of the flower crown. Those pixels are referred to as SF1_B. Finally we transfer the color associated with SF1_B onto F454_B using the transfer algorithm described in the Method section. Result of the transfer is shown in Figure 3.11 E, to be compared with the original, namely the London version shown in Figure 3.11 D. As expected, the sunflowers in the modified London version appear darker than in the original, as they are now put in a context of a darker background.

Restoring colors in van Gogh's *Sunflowers* using color correction. Possibly the most striking difference we observed when comparing the hues of yellow in the flower depiction of the three versions of *Sunflowers in a vase with a yellow background* with the hues of yellow in real sunflowers observed in modern photographs is an increase in the amount of blue coordinates in those yellow hues in the paintings, leading to a less vibrant and brownish yellow, in agreement with chemical analyses of the aging of chrome yellow [**9**, **82**, **83**, **84**, **104**]. This observation hints at an opportunity to restore algorithmically the original colors of the *Sunflowers*: reducing the amount of blue for all pixels in the images of the paintings. We implemented two versions of such an algorithm as follows. We first note that the images of the paintings have been divided into three ROIs, namely the flowers, yf-ROI, the stems of the flowers, g-ROI, and the background, bg-ROI. The first two ROIs are relatively homogeneous in color, yellow and green, respectively, while the latter includes a more diverse spectrum of colors, as it includes the table, vase, and flower background. Each ROI is then processed independently. All pixels within a ROI are clustered first, using the differences in their color as a metric, and hierarchical clustering (see methods). The leaves of the corresponding tree are represented as vertical color lines, where the color of the line is the color of the leaf. Those color lines are organized as a color bar, which is then segmented into clusters. Figures 3.10 A and 3.11 A provide illustrations of such a color bar with its clusters for the yf-ROI and bg-ROI of the

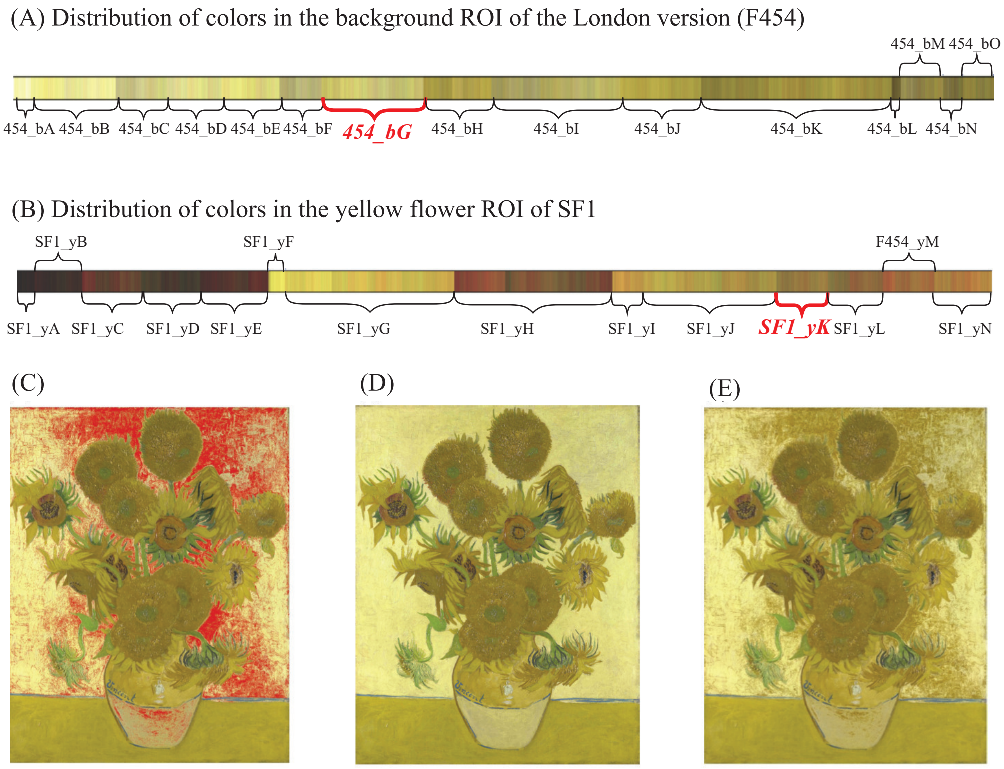

(A) Distribution of colors in the background ROI of the London version (F454)
454_bM 454_bO
454_bA 454_bB 454_bC 454_bD 454_bE 454_bF 454_bG 454_bH 454_bI 454_bJ 454_bK 454_bL 454_bN
(B) Distribution of colors in the yellow flower ROI of SF1
SF1_yB SF1_yF F454_yM
SF1_yA SF1_yC SF1_yD SF1_yE SF1_yG SF1_yH SF1_yI SF1_yJ SF1_yK SF1_yL SF1_yN
(C)
(D)
(E)

FIGURE 3.11. Color transfer from a real sunflower to the yellow part of the background in the London version (A) The color bar representing all hues of colors in the ROI corresponding to the background in the London version. This color corresponds to the leaves of the HC tree computed from the colors of all pixels in this ROI. We select the region (cluster) 454_bG in this color bar which contains some of the pixels that are directly located behind the sunflowers in the vase. The pixels associated to this region are referred to as F454_B. (B) The color bar representing all hues of yellow in the flower ROI of SF1. We identify the cluster SF1_yK in this color bar that contains some of the darkestt yellow. The pixels associated to this region are referred to as SF1_B (C) Pixels in the London version that belong to F454_B are highlighted in red (D) London version: before color transfer (E) London version: after color transfer.

London version, respectively. Each cluster regroups pixels with similar color within an ROI. Those clusters are then processed separately. For a given cluster k within an ROI, we compute first the minimal blue coordinate, m_k, over all pixels in the cluster. Let i be one such pixel, and let b_i be its blue coordinate. We correct this blue coordinate using one of the two following pseudo color restoration (PCR) schemes:

PCR-1: $b_i' = b_i - m_k$

PCR-2: $b_i' = b_i - \min(m_k, 60)$

PCR-1 was designed to remove as much blue as possible from the color coordinates of the pixels in the images of the paintings. This correction is based on the observation that the yellow colors of real sunflowers contain nearly no blue, as seen in Figure 3.6. It is expected to work well for all yellow hues that have been tarnished. PCR-2 is a more gentle correction as it limits the amount of blue that can be deducted. The upper limit of 60 for this deduction comes from the observation that the blue shift detected in the flower regions of the three paintings is close to this value (this is especially clear for the Tokyo version, see Figure 3.5 B).

In Figure 3.12 we show the results of applying the two strategies PCR-1 and PCR-2 described above. There are a few observations associated with those reconstructions. First, visually the differences between PCR-1 and PCR-2 are small. This is expected as those two schemes are expected to act similarly on yellow hues, and yellow dominates in the three paintings. Second, there is a clear difference between the restored images and the original images as the colors appear much brighter in the reconstructed images, especially the yellow in the background (see for example the effects on the London version). While we need to be cautious as to ascertaining that the reconstructions represent the paintings as originally intended by van Gogh, we can safely say that they are most likely closer to his intent, as van Gogh was playing with the contrasts between the sunflowers and the background (as he was experimenting with different colors and intensities for those backgrounds) and that darker colors reduce this contrast (see Figures3.11 D and E). Finally, the reconstructed models of the paintings highlight differences in the contrast between flowers and background between the original, namely the London version, and the two repetitions, the Tokyo and Amsterdam versions, with the original based on a brighter background. This observation reemphasize the importance given by van Gogh to the capture of light, colors, and contrast for the *Sunflowers in a vase*.

3.3. Conclusion

Vincent van Gogh was by no means a chemist by his own admission (letter 889 to his brother Theo [**7**]). However, he was well aware that colors in paintings evolve due to changes in the chemical nature of their pigments: "...paintings fade like flowers" (letter 765 to Theo, [**7**]). In the case of the *Sunflowers in a vase with a yellow background*, namely an original, the London version, and

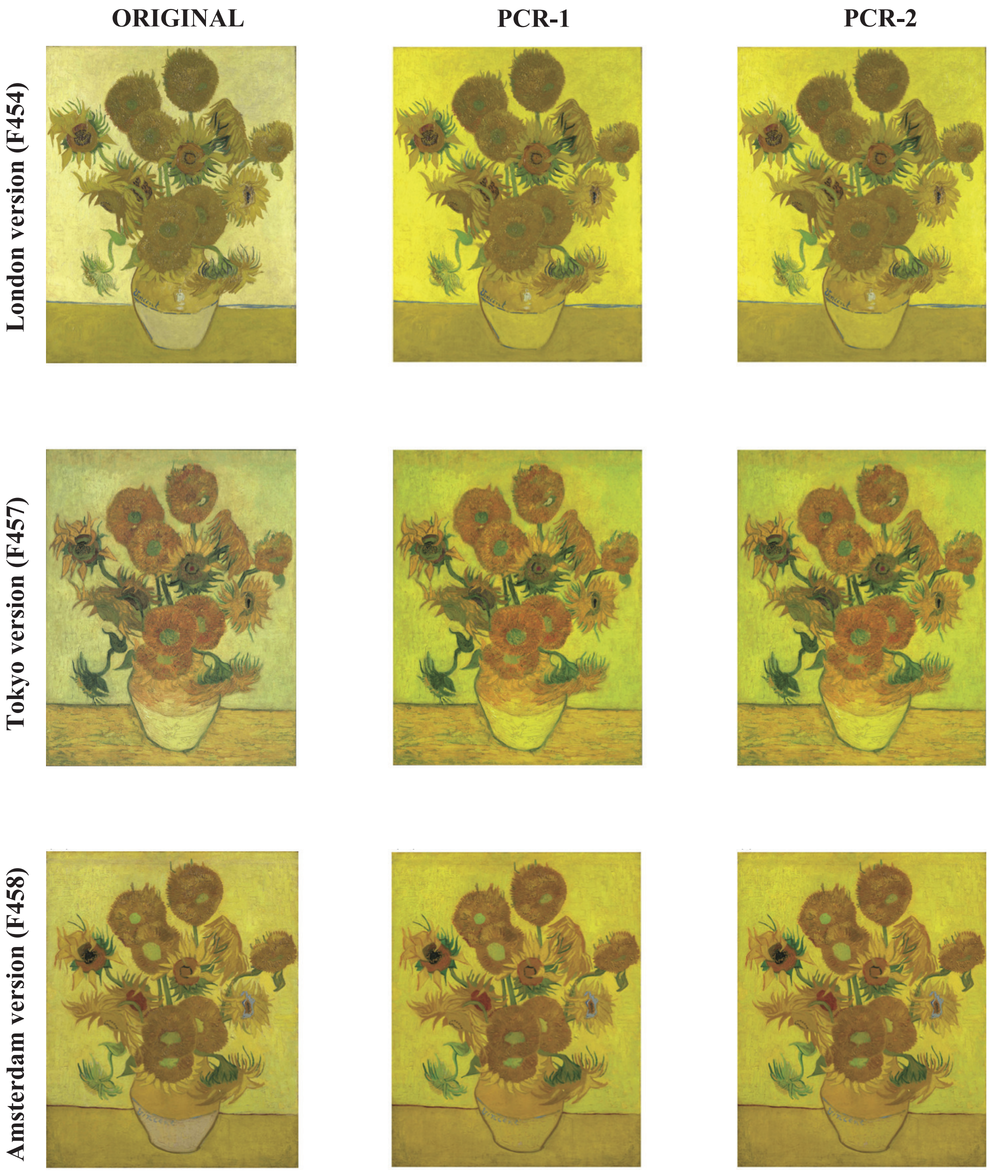


FIGURE 3.12. Color restoration of the *Sunflowers in a vase with a yellow background* For the original sunflower in a vase with yellow background, the London version (top row), and the two repetitions, the Tokyo version and the Amsterdam, bottom row, we show a picture of the current painting (left column), as well as two possible models of the painting intended by van Gogh, using the color reconstruction methods PCR-1 (middle column) and PCR-2 (right column). See text for details.

two replicates, the Tokyo and Amsterdam versions that van Gogh painted using the original as a model, this sentiment came true, as the background and flower rendition in those paintings have faded and turned brown, making them less vibrant than van Gogh had most likely aimed at. His intents with those paintings were to "push chromatic intensity ... with the aim of achieving a radical light-on-light effect" [**8**]. We have attempted to restore van Gogh's intent using a computational approach based on data science. After identifications of regions of interest (ROI) within the three paintings that capture the flowers, stems of the flowers, and background, respectively, we studied the geometry of the color space (in RGB representation) occupied by those ROIs. By comparing those color spaces with those occupied by similar ROIs in photographs of real sunflowers, we identified shifts in all three color coordinates, R, G, and B, with the shift in the blue coordinate being the more salient. This shift in blue that leads to hues of yellow that are faded and even brownish are consistent with the fading of the chrome yellow, the pigments used by van Gogh for representing yellow in *Sunflowers*, observed by chemical spectroscopic methods [**9**, **82**, **83**, **84**, **104**]. We have proposed two algorithms, PCR-1 and PCR-2, for correcting that shift in blue and generate representations of the paintings that aim to restore their original conditions (see Figure 3.12). While we acknowledge that these are models, the reduction of the blue component in the yellow hues has lead to more vibrant and less brownish digital rendition of the three *Sunflowers in a vase with a yellow background*.

While we believe that the models we have generated for the *Sunflowers* take the viewers closer to the paintings created by van Gogh, we acknowledge that the current state-of-the-art techniques for digital reconstructions, including our own, are still limited in number and far from actually restoring the original version of the painter. Progress will ultimately come from combinations of techniques. We did not have access to spectroscopic fluorescence data on the *Sunflowers*; we believe that such data would have helped us delineate the regions of interest in the paintings, in particular those regions where chrome yellow dominates as those regions are more susceptible to fading due to aging of the chrome yellow pigments. In return, the computational methods proposed here for analyzing the RGB space occupied by those regions should help identify the impact in color space of aging, as well as methods for reversing those effects. In addition, those methods enable manipulations of the images of the paintings of interest and therefore the assessment of hypotheses on how the painter was experimenting with color juxtapositions within a painting. Based on our preliminary studies of the *Sunflowers*, it is our intent to develop a general computational framework for digital restoration and manipulation of paintings and make this framework available as a tool for enhancing painting viewing experience.

3.4. Materials and methods

3.4.1. Material: the digital images used in this study. F454 is the fourth version of the *Sunflowers* painted by van Gogh while he was living in Arles, with the specificity of having a yellow background in contrast with the turquoise (F453), royal-blue (F459), and blue-green (F456) for the other versions. It is currently owned by the National Gallery in London, UK, and stored under the inventory number NG3863. The National Gallery only provides a low resolution image of this painting on its website; we found a high resolution version on Wikimedia commons, from the URL `https://upload.wikimedia.org/wikipedia/commons/4/46/Vincent_Willem_van_Gogh_127.jpg`. This high resolution image is provided in jpeg format, with a resolution of $3,349 \times 4,226$ pixels.

F458 is one of the two repetitions of F454 painted by van Gogh, owned by and displayed at the Van Gogh Museum, Amsterdam. The van Gogh museum only provides a low resolution image of this painting on its web site (`https://www.vangoghmuseum.nl/en/collection/s0031V1962`). However, just like for the London version, we found a high resolution version on Wikimedia commons, from the URL `https://upload.wikimedia.org/wikipedia/commons/9/9d/Vincent_van_Gogh_-_Sunflowers_-_VGM_F458.jpg`. This high resolution image is provided in Jpeg format, with a resolution of $3,224 \times 4,226$ pixels.

The high resolution image of F457 was generously provided by the Seiji Togo memorial Sompo Japan Museum of Art, Tokyo, Japan. It was also provided in jpeg format, with a resolution of 3626×4829 pixels.

In addition to the three high resolution images of the three paintings F454, F457, and F458, dubbed the London, Tokyo, and Amsterdam versions, respectively, we used two recent digital images of sunflowers in a field, to capture actual colors of sunflowers. There two pictures, which we label as SF1 and SF2, were obtained from:

SF1) A picture of sunflowers in Provence (Shuttersandsunflowers.com), downloaded with permission of the author.

SF2) A picture of sunflowers in a field, by Leo Adamchuk, available within Adobe Stock and downloaded as part of their free trial.

3.4.2. Color statistics in digital images. A digital image is an image composed of picture elements, or *pixels*. Each pixel is characterized by spatial coordinates, (x, y), that define its positions along the x-axis and y-axis within the image, as well as by color components, (r, g, b), that define the amount of red (R), green (G), and blue (B) which, when combined, describe the color at the pixel. Note that we rely here on the additive RGB color model; this model is not universally accepted and other models such as CMYK (a substractive model) or Lab are possible. We have used RGB as it is the color model that was available with the images we recovered. In this RGB model, each color component r, g, or b is an integer in the range $[0, 255]$. Namely, each fundamental color R, G, or B is described by a discrete variable that can take 256 distinct values, while a composite color is described by its 3 coordinates along those fundamental colors, and therefore belongs to a discrete space (in this case a cube) with $[0, 255]^3$ possible values.

Let us define a digital image I as the set $S(I)$ of its pixels. Let N be the cardinality of that set, namely the total number of pixels in the image. Let C be one of the fundamental colors, namely C can be R, G, or B. As we have seen above, C can take 256 possible discrete values, each in the range $[0, 255]$. We can compute the distributions of these values over a given image as follows. Let $S(c)$ be the set of pixels such that its color coordinate for the color C is c:

$$S(c) = \{p \in P(I) \quad | \quad C(p) = c\}$$

The probability of observing color C with intensity c within the image is then given by:

$$P(C = c) = \frac{|S(c)|}{N} \tag{3.1}$$

where $|S(c)|$ stands for the number of elements of set $S(c)$, i.e., its number of elements. The entropy of the color C within the image is then given by:

$$H(C) = -\sum_{c=0}^{255} P(C = c) \log(P(C = c))$$

The entropy measures the amount of "information" associated with the color C in the image. If the color C is always represented with the same value over each pixel, the entropy is zero, while if the possible values for the color are evenly distributed, the entropy is at its maximum with a value of $\log(256)$.

We can also measure the association of fundamental colors within the image. Let C and D be two of the three fundamental colors. Let $P(C = c, D = d)$ be the joint probability of those colors C and D taking the values c and d, respectively. If we define the set $S(c, d)$ as

$$S(c, d) = \{p \in P(I) \quad | \quad C(p) = c \;\&\; D(p) = d\}$$

then

$$P(C = c, D = d) = \frac{|S(c, d)|}{N} \tag{3.2}$$

From this joint probability, we can compute the conditional probability that $C = c$, knowing that $D = d$, as

$$P(C = c \mid D = d) = \frac{P(C = c, D = d)}{P(d)}$$

.

Let us define now $H(C \mid D = d)$ as the entropy of the color C conditioned on the color D taking the value d. This entropy is computed as:

$$\begin{aligned} H(C \mid D = d) &= -\sum_{c=0}^{255} P(C = c \mid D = d) \log(P(C = c \mid D = d)) \\ &= -\sum_{c=0}^{255} \frac{P(C = c, D = d)}{P(D = d)} \log \frac{P(C = c, D = d)}{P(D = d)} \end{aligned}$$

where the sums extend over all 256 possible values for the coordinates associated to the color C. The entropy of the color C conditioned on the color D is then computed as the average of those numbers over all possible values for d:

$$\begin{aligned} H(C \mid D) &= \sum_{d=0}^{255} P(D=d) H(C \mid D=d) \\ &= -\sum_{d=0}^{255} P(D=d) \sum_{c=0}^{255} \frac{P(C=c, D=d)}{P(D=d)} \log \frac{P(C=c, D=d)}{P(D=d)} \\ &= -\sum_{d=0}^{255} \sum_{c=0}^{255} P(C=c, D=d) \log \frac{P(C=c, D=d)}{P(D=d)} \end{aligned}$$

where the probabilities $P(D=d)$ and $P(C=c, D=d)$ are computed using Equations 3.1 and 3.2, respectively. Finally, we compute the relative conditional entropy as

$$HR(C \mid D) = \frac{H(C \mid D)}{H(C)}$$

$HR(C \mid D)$ takes values between 0 and 1. $HR(C \mid D) = 0$ if and only if the values of C are completely determined by the values of D. Conversely, $HR(C \mid D) = 1$ if and only if the values of C and D are independent of each other.

3.4.3. Hierarchical Clustering Analysis (HCA). Clustering is the task of regrouping objects such that those that belong to the same group, called a cluster, are more similar to each other than to those in other groups. In our analyses, the objects are pixels within an image. A pixel $P(i)$ is characterized by its color, $c(i)$, given as a combination of its composents along the fundamental colors R, G, and B, i.e. $c(i) = (r(i), g(i), b(i))$. As described above, $c(i)$ belongs to the 3D discrete space $[0, 255]^3$. The similarity between two pixels $P(i)$ and $P(j)$ is set to be the Euclidean distance between their colors:

$$d(P(i), P(j)) = \sqrt{(r(i)-r(j))^2 + (g(i)-g(j))^2 + (b(i)-b(j))^2} \tag{3.3}$$

The clustering of the pixels is then performed using the agglomerative hierarchical clustering analysis, or HCA. The is a bottom-up approach in which each pixel starts in its own cluster, and pairs of clusters are merged iteratively until all pixels belong to the same cluster. The whole procedure defines a hierarchy of clusters, also referred as as a clustering tree. A key element to this procedure is to define the distance between two clusters, also referred to as the linkage criterion. When the two clusters contain a single element, this distance is simply the distance between those elements, as defined in equation 3.3. When the two clusters A and B are sets of elements, the distance is then defined as a function of the pairwise distances between those elements. Two common choices are the single linkage:

$$d(A, B) = \min\{d(a, b), a \in A, b \in B\}$$

and the complete linkage:

$$d(A, B) = \max\{d(a, b), a \in A, b \in B\}$$

We have used the complete linkage in all our analyses.

3.4.4. Identifying regions of interest in the paintings *Sunflowers in a vase with a yellow background*. A region of interest (ROI) is defined according to its position in the image as well as on its color consistency. In the *Sunflowers in a vase with a yellow background*, we identify three such ROIs, namely the flower crowns of the sunflowers (yellow), yf-ROI, the stems of the sunflowers (green), g-ROI, and the background, bg-ROI. Automated segmentation is one approach for selecting the pixels that belong to each ROI. We have used instead a combination of manual and automated processing, using the algorithm described below.

This algorithm is based on the fact that the yf-ROI and g-ROI are chosen based on color consistency. We first divide the whole discrete RGB space $[0, 255] \times [0, 255] \times [0, 255]$ into a collection of $5 \times 5 \times 5$ color cubes. All pixels of an image are then attached to those cubes based on their colors. Cubes that are occupied are then assigned a representative, whose color is identified as the center of the cube. All those representatives are then clustered using the HC technique described above, using the Euclidean distance between their colors as a similarity measure, and complete linkage. At the bottom of the HC tree, we draw a color bar with each leave of the tree represented with a vertical line whose color corresponds to the cube associated with that leave. The HC tree is then cut to form 70 clusters; we chose 70 somewhat arbitrarily, such that we would have enough cluster. Most of those clusters have consistent colors along the color bar described above. Two groups of clusters are selected visually, those that are predominantly yellow, with a yellow that resembles the yellow of the flower in the paintings, and those that are predominantly green. Each of this group is then processed separately.

Let G be the group of clusters of green colors identified above. Collect all the pixels in G, and repeat the procedure described above, but with the RGB space divided into a collection of $4 \times 4 \times 4$ cubes. Again, identify all cubes that are not empty, cluster their representatives, and select only those clusters that contain pixels whose colors are consistently green. The remaining pixels are processed one last time, with the RGB space divided into a collection of $3 \times 3 \times 3$ cubes. All the corresponding occupied cubes are ordered with respect to the number of pixels they contain. Cubes are retained by going down on the ordered list, starting with the most populated cube, until 90% of the remaining pixels are selected. Finally, those pixels are identified in the paintings, and those that do not correspond to a flower stem are discarded. The remaining pixels form the green ROI , g-ROI, associated with the stems of the sunflowers.

Let Y be the group of clusters of yellow colors identified in the first step of the algorithm. This group of clusters is processed in the same way as G was processed, leading to a group of pixels that are mostly yellow and that belong to the crowns of the sunflowers. This group of pixels forms divided the yf-ROI, associated with the crowns of the sunflowers.

Finally, all pixels that do not belong to the g-ROI and yf-ROI are deemed to belong to the background (this includes the table in the front, the vase, and the background behind the flowers. Those pixels form the bg-ROI.

3.4.5. Region specific Color transfer between two images. Assume that we have identified a set of pixels A in one picture and that those pixels have similar (also not necessarily equal) colors. Similarly, we have identified a set of pixels B in another picture, with similar colors. A and B may be of different size and may contain different colors. Our goal is to transfer the colors from A to B. Let $A = \{\mathbf{a}_1, \dots, \mathbf{a}_i, \dots, \mathbf{a}_N\}$ and $B = \{\mathbf{b}_1, \dots, \mathbf{b}_j, \dots, \mathbf{b}_M\}$ where the $\mathbf{a}$s and $\mathbf{b}$s are 3D vectors containing the R, G, and B coordinates of the colors of the pixels in A and B, respectively. We apply the following algorithm:

i) Compute the centers of mass of A and B: $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$;

ii) Translate B to B^* so that A and B' have the same center of mass: $B^* = \{\mathbf{b}_1^*, \dots, \mathbf{b}_j^*, \dots, \mathbf{b}_M^*\}$ with $\mathbf{b}_j^* = \mathbf{b}_j - \bar{\mathbf{b}} + \bar{\mathbf{a}}$;

iii) Build a HC-tree on the union of A and B^*, using the Euclidean distance as a metric and complete linkage, and cut at a tree level to define clusters. If a cluster only contains elements originally from B^*, merge it with its closest cluster (in terms of tree distance) that contains elements of A. At the end of this procedure, we have a set of clusters, with each cluster containing either elements of A and B^*, or elements of A only. The latter are discarded.

iv) Let C be one of the clusters from step iii), and let $AC = \{\mathbf{a}_1, \ldots, \mathbf{a}_i, \ldots, \mathbf{a}_{NC}\}$ and $BC = \{\mathbf{b}_1^*, \ldots, \mathbf{b}_j^*, \ldots, \mathbf{b}_{MC}^*\}$ be the subsets of A and B^* that belong C, where NC and MC are their sizes, respectively. For each j in $[1, MC]$, replace $\mathbf{b}_j^*$ with a color picked randomly in AC. Repeat over all clusters C.

At the end of this procedure, each element j of B has been assigned a new color b_j^* that is inherited from A. The pixels in the second picture are then assigned those new colors.

CHAPTER 4

Unraveling heterogeneity of ADNI's time-to-event data using conditional entropy Part-I: Cross-sectional study

4.1. Introduction

Besides the pressing global climate issues, our humanity is facing another pressing issue: human aging. On October 1st, 2022 World Health Organization (WHO) released news regarding populations of elderly in the world quoted as:
" By 2030, 1 in 6 people in the world will be aged 60 years or over. At this time the share of the population aged 60 years and over will increase from 1 billion in 2020 to 1.4 billion. By 2050, the world's population of people aged 60 years and older will double (2.1 billion). The number of persons aged 80 years or older is expected to triple between 2020 and 2050 to reach 426 million." To be more specific, a mentioned example in the news release is:"Japan 30% of the population is already over 60 years old." That is, any common health issue related to aging would be one on a global scale.

In particular, as population aging is coming fast, so is human brain aging [**21**,**92**]. Here, brain aging primarily means changes in cognitive functions [**40**]. When such changes go toward the dad directions, such as Alzheimer's disease as the focus of this study among many other diseases, the pressing health issue of human aging zooms really big.

As quoted from the home page of the Alzheimer's Disease Neuroimaging Initiative (ADNI) website: (https://adni.loni.usc.edu),
"Alzheimer's disease (AD) is an irreversible neurodegenerative disease that results in a loss of mental function caused by the deterioration of brain tissue. It is the most common cause of dementia among people over the age of 65, affecting an estimated 5.5 million Americans, yet no prevention methods or cures have been discovered. For more information about Alzheimer's disease, visit the Alzheimer's Association."

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease (AD). Since its launch more than a decade ago, the landmark public-private partnership has made major contributions to AD research, enabling the sharing of data between researchers around the world.

The goal of this study is in particular aligning with ADNI's first overarching goal: To detect AD at the earliest possible stage (pre-dementia) and identify ways to track the disease's progression with biomarkers. See the detailed Goals and Study design of the ADNI website at (http://www.adni-info.org.).

So far studies using ADNI cross-sectional and longitudinal data from multiple modalities have reported two particularly relevant pieces of information: 1) AD pathology is already present in people with no outward sign of memory loss and these cognitively normal people may already have subtle brain atrophy; 2) Both the Cognitively Normal (CN) and Mild Cognitive Impairment (MCI) groups are pathologically heterogeneous. Some people show no signs of AD, some show signs of progressing to AD quickly, and others show signs of progressing to dementias other than AD.

Many cross-sectional studies on tracking progressions of AD with biomarkers using data collected from the ADNI database have reported applying the well-known Cox Proportional Hazard Regression model for the time-to-event data: from CN to MCI or from MCI to AD, and its well-studied partial likelihood approach as studies' chief inferential apparatus for selecting so-called significant biomarkers. However, this popular methodology of Survival analysis in Statistics has to face a series of fundamental questions when dealing with real-world data. When analyzing data from the ADNI database, not surprisingly, these essential questions, as listed below, are left unanswered in all studies found in literature [**61**, **66**, **69**, **86**]:

Q1:: Does the Cox PH modeling assumptions on non-informative censoring mechanism violate the pattern-information embedded within data?

Q2:: Does the data support the linearity-based additive effects of covariate features assumed by the Cox PH modeling structure?

Q3:: Given that heterogeneity among subjects is intrinsic in ADNI data, could Cox PH provide reliable results, in particular when facing heavy censoring rates on the global and local scales and potentially complex interacting relations and effects among covariate features?

Q4:: Is the partial likelihood based inferential approach valid?

Q5:: If both the Cox PH model and the partial likelihood approach fail fundamentally, what are potential resolutions to achieve the overreaching goal of ADNI in this complex disease AD?

Q6:: After all, how do we compute and identify pertinent perspectives of heterogeneity in cross-sectional ADNI data?

Q7:: With the computed presence of heterogeneity, how to display the information contained in full?

We address Q1 through Q5 to a great extent, but only partially touch on Q6 and Q7 in this paper. Detailed and full discussions of the last two questions are deferred to Part-II. Here, we briefly explain why this series of questions is critical from a scientific perspective.

Testing whether the censoring scheme is non-informative or not in Q1 is the starter of data analysis applying methodologies in Survival Analysis, such as Cox PH modeling. However, this testing has not been carried out in the ADNI-associated literature. The reason might be that there is no simple testing statistic available in the Survival Analysis literature. As for Q2, the linearity-based additivity of covariate effects obviously doesn't work for categorical covariate variables, which is a common data type, such as sex and education. Even for quantitative variables, the sum of measurements of very distinct metric units is rather hard to explain literately and convincingly. With the presence of heterogeneity among subjects, Cox PH becomes rather limited with respect to fronts: censoring rates and unknown functional forms of interaction. The overall censoring rate is already high. It is more than 60% in this cross-sectional study. The censoring rates can reach more than 90% in some sub-collections of subjects defined by the potential perspective of heterogeneity. On the other front, as multiple reports depicting interacting effects in AD literature [**4**, **5**, **25**], such real-world interacting relations among covariate variables might be known prior, but their effects are hardly known functionally. Thus, any functional forms of interacting effects among covariate variables deem unrealistic and dangerous to subject matter science because of misinformation.

For Q4: the partial likelihood approach is strictly based on the correlation for evaluating associative relations between the time-to-event response variable and covariate features. This fact can be easily seen from the score equations derived from the partial likelihood, which involves $T_{(i)}$ minus its conditional expectation within the risk set. This format of partial likelihood score equations points to one strict and fundamental associative concept: correlation. It is known that the correlation concept is strictly designed for one quantitative variable against another quantitative variable, not for categorical ones. It is not valid for evaluating 1-to-2 or 2-to-k associative relations. This limit is consequential for the following reason. When facing an informative censoring scheme, the response and censoring variables are dependent. Then, the proper response variable must be 2D bivariate (T_i, C_i). Under such a setting, the partial likelihood approach would not work. This fact again points to the critical role of Q1.

For Q5 and partially for Q6 and Q7, this paper proposes the Categorical Exploratory Data Analysis (CEDA) paradigm to resolve all issues raised from Q1 through Q4. In CEDA, we apply Shannon conditional entropy to evaluate associative relation, and mutual information to select so-called major factors underlying the dynamics of the response variable in relation to all covariate feature variables. When using two key Theoretical Information measurements, the CEDA naturally adopts the contingency table as its basic computational platform. Consequently, CEDA not only works for all data types, including the categorical one but also is capable of evaluating $k - to - k'$ associative relations for all integers k and k'. Since k categorized or categorical variable can be fused into one categorical variable via their contingency k-dim hypercubes. When facing censored data disregarding the censoring rate, we apply the re-distribution-to-the-right algorithm [**26**] to build all sorts of contingency tables. For this simple reason, CEDA is capable of handling even 100% censoring rates in any locality. We also can build the contingency table for categorized T_i and C_i to resolve the Q1 formally. This is a new resolution. And because CEDA is able to identify major factors of varying orders, the task of identifying pertinent perspectives of heterogeneity in Q6 is resolved to a great extent. We also develop a display called conditional-entropy-expansion to unravel the effects of all chosen perspectives of heterogeneity.

The organization of this paper is given as follows. In Section 4.2, we describe the cross-sectional ADNI data by providing brief descriptions for all variables: response, censoring, and 16 covariate features. In Section 4.3, we first review Theoretical Information measurements and rationales for major factor selection under independent and dependent covariate settings and illustrate how the Redistribution-to-the-right algorithm precisely works. In Section 4.4, we conduct three simulated experiments with varying censoring rates to showcase information relevant to Q1-Q5. The ADNI data is thoroughly analyzed in the lengthy Section 4.5 with two perspectives of heterogeneity being studied and presented. In the last section of the Conclusion, we discuss the contributions of this Part-I paper, and briefly lay out computational tasks in Part-II. We hope this paper would help scientists to advance their research on Alzheimer's Disease and we are thankful to ADNI for making such an important database available to us.

4.2. ADNI Data Description

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

We accessed the data file on June 9, 2022. It included 15,941 records from 1,094 participants. By removing any subjects with missing data, we ended up with the records of 903 subjects with a baseline diagnosis of Mild Cognitive Impairment (MCI) and at least one follow-up diagnosis. There are 346 subjects (211 males; 135 females) whose diagnosis progressed to Dementia (AD) and their time-to-even T_i's observed. During the follow-up, 557 individuals (328 males; 229 females) remained MCI, and their last scheduled exam times are observed as censoring time C_is. Therefore, we observe $Y_i = (T_i \wedge C_i)$, the minimum of T_i and C_i, and the censoring status $\delta_i = 1_{[T_i \leq C_i]}$ for each of these 903 subjects.

Data feature descriptions are given as follows. The 16 covariate features used here are coming from three categories: 4 demographic features ($V1 - V4$), 1 MRI-related feature ($V5$), and 11 clinical features ($V6 - V16$). The features' abbreviation and definition are provided in Table 4.1 followed by detailed descriptions of 11 clinical features.

Index	**features**	**Definition**
$V1$	age	Age at baseline
$V2$	gender	Gender
$V3$	Education	Education Level
$V4$	APOE $\epsilon 4$	Apolipoprotein E $\epsilon 4$ allele
$V5$	FLDSTRENG-bl	MRI Scanner's Field Strength (1.5T or 3 T) used at baseline
$V6$	CDR-SB-bl	Clinical Dementia Rating at baseline (the sum of boxes)
$V7$	FAQ	Functional Activities Questionnaire
$V8$	ADAS13-bl	13-item AD Assessment Scale-Cognitive Subscale at baseline
$V9$	MEM-mean	Mean of the Composite Cognitive Score for Memory
$V10$	MEM-std	"std" of the Composite Cognitive Score for Memory
$V11$	EXF-mean	Mean of the Composite Cognitive Score for Executive Functioning
$V12$	EXF-std	"std" of the Composite Cognitive Score for Executive Functioning
$V13$	LAN-mean	Mean of the Composite Cognitive Score for Language
$V14$	LAN-std	"std" of the Composite Cognitive Score for Language
$V15$	VSP-mean	Mean of the Composite Cognitive Score for Visuospatial Functioning
$V16$	VSP-std	"std" of the Composite Cognitive Score for Visuospatial Functioning

TABLE 4.1. Covariate features as biomarkers in the ADNI data analysis. "std": the standard deviation.

Some further pieces of information regarding three clinical features: CDR-SB (V6) [**85**] indicates the sum of scores for the following six domains of functioning: memory, orientation, judgment and problem solving, community affairs, home, and hobbies, and personal care. The CDR-SB ranges from 0 (no impairment) to 18 (severe impairment in all six domains); FAQ (V7) measures activities of daily living and ranges from 0 to 30 with higher scores reflecting greater cognitive impairment [**89**]; The modified ADAS-Cog 13-item scale (ADAS13(V8)) [**79**] is a modified evaluation by adding a number cancellation task and a delayed free recall task to the 11-item ADAS-Cog (ADAS11). The higher scores suggest greater impairment.

In addition to the clinical features aforementioned, we consider four composite cognitive scores of 4 domains, including memory, executive functioning, language, and visuospatial functioning [**87**]. The scores are developed and calibrated by using the ADNI Neuropsychological batteries. During each available follow-up, the four composite cognitive scores are updated accordingly and thus they are time-varying. To extract information and characteristics for consideration, we make use of the mean and standard deviation of the composite cognitive score for each category until the event (i.e., AD or the last time point if censored) as two new features to describe the overall characteristics of each score. Therefore, we have eight total new features to represent the four composite cognitive scores which are shown in Table 4.1 ($V9 - V16$). With such construction of time-varying scores of these 4 domains, we expect each subject has the mean and standard deviation of the composite cognitive scores at every exam time. Missing values of the composite cognitive scores at partial exam time points are found due to the intrinsic incomplete data (e.g., participant refusal to complete the study or the item was not administered due to the time limit) and this can be fixed by taking the average of available scores correspondingly. We remove subjects with no scores available across all exam time points as there is not enough useful information provided.

4.3. CEDA Methodologies

In this section, we first briefly review computational developments for CEDA's major factor selection protocol. This protocol is entirely based on Theoretical Information Measurements: marginal and conditional entropies and mutual information. Since we work only with categorical or categorized variables in CEDA. Therefore, the only version of entropy used here is Shannon entropy. Secondly, we illustrate the "Redistribution-to-the right" algorithm for building a contingency table with the presence of the right censored data. Since all CEDA computing is performed upon the contingency table platform. Thus, this algorithm indeed plays a critical role in this paper.

4.3.1. CEDA's major factor selection protocol. In this subsection, we briefly review the concept and computing of conditional entropy and mutual information as two key Theoretical Information Measurements used throughout this paper. Detailed derivations of related formulas of these two measurements are referred to previous works in [**14, 15, 17, 36**]. We employ these two entropy-based measurements to evaluate potentially nonlinear directional association from a generic covariate feature variable denoted as X to a generic response variable denoted as $\mathcal{Y}$, and the nondirectional association between two covariate feature variables, say X_1 and X_2.

Since we exclusively work on categorical or categorized variables in this paper. That is, any variable of continuous or discrete measurements is categorized with respect to a chosen version of the variable's histogram. For instance, the time-to-event T, the mean and standard deviation of test scores, and age are to be categorized. Further, any set of multiple categorical variables can be fused into a new categorical variable by redefining each distinct multiple-dimensional categorical vector as a category of the newly defined categorical variable. For instance, a bivariate vector (X_1, X_2) would be taken and treated as a categorical variable. That is, the categorical variable, such as gender (V2), and the categorized variable, such as MEM-mean (V9), can be fused into a new categorical variable. Therefore, any set of covariate feature categorical or categorized variables, say A, is also taken as a categorical variable with the same name A. So A's directional association to $\mathcal{Y}$ is evaluated in the same way as evaluating the directional association of any member of A to a categorical response variable $\mathcal{Y}$.

Consider A or B as two different categorical covariate variables standing for two sets of covariate features. We evaluate the directional association of A to $\mathcal{Y}$ upon their contingency table $C[A - vs - \mathcal{Y}]$ with categories of A and $\mathcal{Y}$ being arranged along the row- and column-axes, respectively, as a conventional format used throughout this paper. Along the row-axis, each row of cell counts in $C[\mathcal{Y} - vs - A]$ is taken to define a conditional multinomial random variable, which is specified by its row-sum and the row-vector of proportions. For instance, a conditional (Shannon) entropy (CE) of $\mathcal{Y}$ at the a-th row of $C[A - vs - \mathcal{Y}]$ is calculated and denoted as:

$$H[\mathcal{Y}|A = a] = (-1) \sum_{y \in \{y_1,..,y_r\}} \hat{p}[\mathcal{Y} = y|A = a] \log \hat{p}[\mathcal{Y} = y|A = a],$$

with $(\hat{p}[\mathcal{Y} = y_1|A = a], .., \hat{p}[\mathcal{Y} = y_r|A = a])$ as the a-th row's vector of proportions. This quantity of $H[\mathcal{Y}|A = a]$ indicates the amount of uncertainty about $\mathcal{Y}$ given the information of $A = a$ being known.

In contrast, the overall amount of uncertainty about $\mathcal{Y}$ given the information of A with $a \in \{a_1, .., a_h\}$ is evaluated as the weighted average and denoted as:

$$H[\mathcal{Y}|A] = \sum_{a \in \{a_1,..,a_h\}} \frac{n_a}{n} H[\mathcal{Y}|A = a],$$

with n_a being a-th row sum and the total sample size $n = \sum_{a \in \{a_1,..,a_h\}} n_a$. Further, the entropies of marginal column-wise vector of proportions $(\frac{n_{y_1}}{n}, ..., \frac{n_{y_r}}{n})$ and row-wise vector of proportions $(\frac{n_{a_1}}{n}, ..., \frac{n_{a_h}}{n})$ are denoted as $H[\mathcal{Y}]$ and $H[A]$, respectively.

It is known that the conditional entropy (CE) $H[\mathcal{Y}|A]$ conveys the expected amount of remaining uncertainty in $\mathcal{Y}$ after knowing A. Likewise, $H[A|\mathcal{Y}]$ conveys the expected amount of remaining uncertainty in A after seeing $\mathcal{Y}$. The two conditional entropy drops, i.e. differences $H[Y] - H[\mathcal{Y}|A]$ and $H[A] - H[A|\mathcal{Y}]$, indicate the shared amount information between A and $\mathcal{Y}$:

$$\begin{aligned} H[\mathcal{Y}] - H[\mathcal{Y}|A] &= H[A] - H[A|\mathcal{Y}] \\ &= H[A] + H[Y] - H[A, \mathcal{Y}] \\ &= I[\mathcal{Y}; A]. \end{aligned}$$

where $I[\mathcal{Y}; A]$ denotes the mutual information between $\mathcal{Y}$ and A.

Further, the conditional mutual information between the bivariate variable (A, B) given $\mathcal{Y}$ is evaluated as:

$$I[A; B|\mathcal{Y}] = H[A|\mathcal{Y}] + H[B|\mathcal{Y}] - H[(A, B)|\mathcal{Y}].$$

Therefore, the mutual information $I[\mathcal{Y}; (A, B)]$ can be estimated and decomposed as follows:

$$\begin{aligned} H[\mathcal{Y}] - H[\mathcal{Y}|(A, B)] &= H[(A, B)] - H[(A, B)|\mathcal{Y}]; \\ &= H[A] + H[B] - I[A; B] - \{H[A|\mathcal{Y}] + H[B|\mathcal{Y}] - I[A; B|\mathcal{Y}]\}; \\ &= \{H[\mathcal{Y}] - H[\mathcal{Y}|A] + H[\mathcal{Y}] - H[\mathcal{Y}|B]\} + \{I[A; B|\mathcal{Y}] - I[A; B]\}, \end{aligned}$$

where the first two terms are individual CE-drops attributed to A and B and the third term is the difference of conditional and marginal mutual information of A and B. In particular, we term A and B achieve their ecological effect if this term: $\{I[A; B|\mathcal{Y}] - I[A; B]\}$, is positive. This positiveness indicates the potential for A and B being concurrently present within the dynamics underlying $\mathcal{Y}$. The essence of achieving the ecological effect is that A and B have the potential of being conditional dependent under $\mathcal{Y}$ disregarding whether they are marginal dependent or not.

However, the above decomposition precisely conveys the interpretable meaning of conditional mutual information when $I[A; B] = 0$ as the two involving feature sets A and B are marginally independent. Thus, if $I[A; B|\mathcal{Y}]$ is significantly larger than $\min\{H[\mathcal{Y}] - H[\mathcal{Y}|A], H[\mathcal{Y}] - H[\mathcal{Y}|B]\}$, then we are certain that A and B achieve a significant interacting effect in reducing the uncertainty of $\mathcal{Y}$. Therefore, we particularly evaluate the so-called successive conditional entropy (SCE) drop as:

$$H[\mathcal{Y}] - H[\mathcal{Y}|(A,B)] - \max\{H[\mathcal{Y}] - H[\mathcal{Y}|A], H[\mathcal{Y}] - H[\mathcal{Y}|B]\}.$$

The task of identifying any realistic interacting effect is always essential in any real-world data analysis because such effects could provide a critical understanding of the system under study. Thus, this SCE-drop would be reported in all tables. Its merit also includes checking whether A and B achieve their ecological effect, which is required before considering whether they have an interacting effect or not.

On the other hand, if A and B are indeed highly associated in the marginal sense via certain unknown dependency, then $I[A;B] > 0$. That is, the term $\{I[A;B|\mathcal{Y}] - I[A;B]\}$, could be negative. Hence, when A and B are associated, we face two chief difficulties. The first difficulty is that it is hard to determine whether the minimum CE-drop: $\min\{H[\mathcal{Y}] - H[\mathcal{Y}|A], H[\mathcal{Y}] - H[\mathcal{Y}|B]\}$, due to either A or B is indeed significant or not. That is since their ecological effect is failed to be seen, we can not be sure whether A and B are concurrently present within the dynamics underlying $\mathcal{Y}$. The second difficulty is that, even $\{I[A;B|\mathcal{Y}] - I[A;B]\}$ is positive with a moderate, not large enough size, then it becomes difficult to assess whether the A and B have a significant interacting effect or not.

In order to resolve these two difficulties, a de-associating procedure is proposed in [**36**] by simply subdividing the entire data set with respect to a target covariate variable's categories. For instance, A is the target covariate variable, and the entire data set is divided into h sub-collections with respect to $\{a_1, .., a_h\}$. That is, A is a constant within each sub-collection. Hence, the association between B and A within each of these h sub-collection disappears. Overall speaking, all covariate variables are less associated within each sub-collection. The merit of the de-associating procedure is evidently seen in Section 4.5 of ADNI data analysis. In fact, the most critical merit of this de-associating procedure indeed rests on the fact that A's perspective of heterogeneity embedded within the dynamics underlying $\mathcal{Y}$ can be much more easily discovered. As such we can discover two versions of heterogeneity from the perspectives of A and B and compare these two versions of heterogeneity. By doing so, we discover authentic heterogeneity-based pattern information hidden in data. Ideally, if we could identify all relevant perspectives of heterogeneity, then collectively we can have the data's full information content. This is termed the ideal scenario in data analysis.

By summarizing all aforementioned developments in this subsection, we arrive at the CEDA's Major Factor Selection (MFS) protocol under two settings. The first setting is developed in [**14**] to deal with independent or slightly dependent covariate feature variables only, while the second set is developed in [**36**] to deal with the heavily dependent covariate setting. Nonetheless, by adopting the de-associating procedure within any heavily dependent setting, the MFS protocol originally built under an independent setting is indeed applied within the sub-collection levels, at which all heavily dependent covariate feature indeed become much less dependent. As such, we discover which covariate features or feature-sets can indeed provide extra amounts of information beyond which targeted covariate features or feature-sets. This computational capability is essential in any real data analysis. Since the goal of data analysis is aimed at the ideal scenario: data's full information content.

It is worth reiterating that, with the potential presence of heterogeneity, the goal of MFS protocol is not set to create one ultimate collection of major factors of various orders that can collectively and concurrently reduce the uncertainty of $\mathcal{Y}$ to the lowest level. The goal of the MFS protocol is to precisely explore and extract pattern information pertaining to perspective-specific heterogeneity. By exploring many perspectives of heterogeneity, we hope we can get much closer to the data's full information content.

4.3.2. Redistribution-to-the-right. Evaluations of association between two categorical or categorized variables are performed on the contingency table platform. Without involving censored data points, the construction of a contingency table is straightforward by counting the number of data points falling into each cell defined by one category from each of the two variables. Nonetheless, when censoring is involved with one variable such as the response survival time variable T, then constructing a contingency table of any covariate variable X against T is not a straightforward computing task. We illustrate how to achieve such a task in this subsection.

Denote the contingency table to be built as $C[X - vs - T]$ with T being censored by C. In this subsection, we first consider the variables T and C being measured at their original time scale before being categorized. Recall that the possibly right-censored survival time data set is denoted as $\{(Y_i, \delta_i, \mathcal{X}_i | i = 1, .., n\}$ with $Y_i = (T_i \wedge C_i)$ as the minimum of T_i and C_i and $\delta_i = 1_{[T_i \leq C_i]}$ indicating the binary censoring status: 0 for being censored and 1 for being uncensored. Let $n_c = n - \sum_{i=1}^{n} \delta_i = n - n_u$ denote the number of censored data points and n_c/n the censoring rate of this data set. For simplicity, we assume all $\{Y_i\}$ are distinct and its order statistics is denoted as $\{Y_{(i)}\}$. The K-dim covariate vector $\mathcal{X}_i = (X_{1i}, X_{2i}, .., X_{Ki})$ records the measurements of K feature variables $\{X_k | k = 1, .., K\}$.

The Kaplan-Meier estimation [**57**] of survival function $S(t) = Pr[T_i > t]$ of T_i based only on $\{(Y_i, \delta_i) | i = 1, .., n\}$, as if without knowledge of $\{\mathcal{X}_i | i = 1, .., n\}$, is built as [**77**]:

$$\hat{S}(t) = \prod_{i:Y_{(i)} \leq t} (1 - \frac{1}{n - i + 1})^{\delta_{(i)}}.$$

It is evident that this empirical distribution $\hat{S}(t)$ has all its jumps at uncensored time points $\{(Y_i, \delta_i) | \delta_i = 1\}$ with possibly unequal jump-sizes. This phenomenon is characterized as so-called "redistribution-to-the-right" in [**26**], that is, the empirical weight $\frac{1}{n}$ of any censored data point is equally redistributed to all data points: uncensored as well as censored, found on its right-hand side. A weight received by any censored data point is likewise re-distributed to all data points on its right. In Table 4.2, we specifically illustrate this redistributing algorithm for the 3rd ordered survival time $Y_{(3)}$, which is censored ($\delta_{(3)} = 0$) among $10(= n)$ data points with the 6th and 8th being censored as well. For expositional simplicity, we denote and mark these three ordered censored time points as $Y^c_{(3)}$, $Y^c_{(6)}$, and $Y^c_{(8)}$.

It is noted that not only $Y^c_{(3)}$ is subject to the such redistribution-the-right algorithm in constructing the empirical survival distribution $\hat{S}(t)$, but also all covariate measurements $\{X_{k(3)} | k = 1, .., K\}$ are subject to the same redistribution when we construct contingency table $C[X_k - vs - T]$. It is because that we only observe $(X_{k(3)}, Y^c_{(3)})$, not the unobserved $(X_{k(3)}, T^c_{(3)})$. And the unobserved $T^c_{(3)}$ is indeed larger than $Y^c_{(3)}$. Since there are two more censored survival time points beyond $Y^c_{(3)}$. The entire redistribution algorithm takes three steps to finish. In Table 4.3, we show the results of the redistribution of empirical weights pertaining to the three censored data points.

Upon the entire data set $\{(Y_i, \delta_i, \mathcal{X}_i | i = 1, .., n\}$, a $n \times n$ weight-redistribution matrix is constructed and fixed according to the layout of n order statistics' $\{(Y_{(i)}, \delta_{(i)})\}$ censoring statuses. Denote this weight-redistribution matrix as $\mathcal{W}$ as shown in Table 4.3. It is noted that

X/Y	$Y_{(1)}$	$Y_{(2)}$	$Y^c_{(3)}$	$Y_{(4)}$	$Y_{(5)}$	$Y^c_{(6)}$	$Y_{(7)}$	$Y^c_{(8)}$	$Y_{(9)}$	$Y_{(10)}$
X_{k3}	0.0	0.0	0.0	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$
X_{k3}	0.0	0.0	0.0	$\frac{1}{7}$	$\frac{1}{7}$	0.0	$\frac{1}{7}+\frac{1}{28}$	$\frac{1}{7}+\frac{1}{28}$	$\frac{1}{7}+\frac{1}{28}$	$\frac{1}{7}+\frac{1}{28}$
X_{k3}	0.0	0.0	0.0	$\frac{1}{7}$	$\frac{1}{7}$	0.0	$\frac{1}{7}+\frac{1}{28}$	0.0	$\frac{1}{7}+\frac{1}{28}+\frac{5}{56}$	$\frac{1}{7}+\frac{1}{28}+\frac{5}{56}$

Table 4.2. A censored data point's (X_3) step-by-step redistribution-to-the-right at three ordered censored time points $Y^c_{(3)}$, $Y^c_{(6)}$ and $Y^c_{(8)}$ along the axis. All row-sums of weights are equal to 1.

X/Y	$Y_{(1)}$	$Y_{(2)}$	$Y^c_{(3)}$	$Y_{(4)}$	$Y_{(5)}$	$Y^c_{(6)}$	$Y_{(7)}$	$Y^c_{(8)}$	$Y_{(9)}$	$Y_{(10)}$
$X_{k(1)}$	1	0	0	0	0	0	0	0	0	0
$X_{k(2)}$	0	1	0	0	0	0	0	0	0	0
$X_{k(3)}$	0	0	0	$\frac{1}{7}$	$\frac{1}{7}$	0	$\frac{5}{28}$	0	$\frac{15}{56}$	$\frac{15}{56}$
$X_{k(4)}$	0	0	0	1	0	0	0	0	0	0
$X_{k(5)}$	0	0	0	0	1	0	0	0	0	0
$X_{k(6)}$	0	0	0	0	0	0	$\frac{1}{4}$	0	$\frac{3}{8}$	$\frac{3}{8}$
$X_{k(7)}$	0	0	0	0	0	0	1	0	0	0
$X_{k(8)}$	0	0	0	0	0	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$
$X_{k(9)}$	0	0	0	0	0	0	0	0	1	0
$X_{k(10)}$	0	0	0	0	0	0	0	0	0	1

Table 4.3. Redistribution-to-the-right of three censored data data points among 7 uncensored ones along the ordered survival time axis. All row-sums of weights are equal to 1.

all columns according to all censored survival times $\{Y_{(i)}|\delta_{(i)} = 0\}$ are n-dim zero-vectors in $\mathcal{W}$. Thus, when the uncensored survival times are categorized by a specific way of grouping on $\{T_{(i')}|\delta_{(i')} = 1, i' = 1, .., n_u\}$, then $\mathcal{W}$ will be subject to the same grouping along its column axis. Consequently, the weights contributed to each bin by any k-th covariate measurement of i-th individual is specified. In this fashion, we are able to construct contingency tables $C[X_k - vs - T]$ for evaluating associations between X_k and T. We can likewise construct a contingency table $C[(X_{k_1}, .., X_{k_l}) - vs - T]$ for a feature-set $\{X_{k_1}, .., X_{k_l}\}$ and T.

4.4. Computer experiments with increasing right censoring rates.

In this section, we report a simple computational study by applying methodologies proposed in the previous section on three simulated right-censored survival time data sets. These three data sets are generated by the same functional structures but have distinct censoring rates: 10%, 20%, and 30%. We hope to demonstrate the stably evolving conditional entropy evaluations and at the same time to show correct selections of major factors across these three different censoring rates.

We employ an integral equation, which has been proposed and studied in [**35**], to generate survival time T as the time of using up the unobserved reserve value U with respect to an exhausting rate specified by $e^{\{V1+sin(2\pi(V2+V3))+V7^2\}}\lambda_0(t)$. The term $\lambda_0(t)$ is taken as the baseline hazard rate. The integral equation is given as follows:

$$U = \int_0^T e^{\{V1+sin(2\pi(V2+V3))+V7^2\}}\lambda_0(t)dt.$$

Here the term $sin(2\pi(V2+V3))$ in the exponent of the integrand is designed to have an interacting relational effect of variables V2 and V3 through a sine function. This simple functional form signals the nonlinearity, on one hand, and the departure from the classic product format of interacting effect, on the other.

Denote the hazard rates of U and T as $\lambda_U(\cdot) = \Lambda'_U(\cdot)$ and $\lambda_T(\cdot) = \Lambda'_T(\cdot)$, respectively. Their relationships are characterized as follows, see more details in [**35**]:

$$\begin{aligned}
& e^{-\Lambda_T(t)} = Pr[T > t] \\
= \quad & Pr[\int_0^T e^{\{V1+sin(2\pi(V2+V3))+V7^2\}}\lambda_0(s)dts > \int_0^t e^{\{V1+sin(2\pi(V2+V3))+V7^2\}}\lambda_0(s)ds] \\
= \quad & Pr[U > \int_0^t e^{\{V1+sin(2\pi(V2+V3))+V7^2\}}\lambda_0(s)ds]; \\
= \quad & e^{-\Lambda_U(\int_0^t e^{\{V1+sin(2\pi(V2+V3))+V7^2\}}\lambda_0(s)ds)}.
\end{aligned}$$

Therefore, we have the cumulative hazard rate and hazard rate of T being specified as follows:

$$\begin{aligned}
\Lambda_T(t) &= \Lambda_U(\int_0^t e^{\{V1+sin(2\pi(V2+V3))+V7^2\}}\lambda_0(s)ds); \\
\lambda_T(t) &= \lambda_U(\int_0^t e^{\{V1+sin(2\pi(V2+V3))+V7^2\}}\lambda_0(s)ds) \\
&\cdot \quad e^{\{V1+sin(2\pi(V2+V3))+V7^2\}}\lambda_0(t).
\end{aligned}$$

That is, if $\lambda_U(\cdot)$ is a constant function, that is, $U\ Exp(.)$, then we have:

$$\lambda_T(t) = e^{\{V1+sin(2\pi(V2+V3))+V7^2\}}\lambda_0(t),$$

which is in a format of Cox's proportional hazard setting.

To simulate three experimental data sets, for simplicity, we use the Weibull baseline hazard function $\lambda_0(t) = kt^{k-1}$ with $k = 1.5$ and Exponential distributed reserve function, $U \sim Exp(1.5)$. There are 10 mutually independent covariate variables $\{V1, .., V10\}$. They are all randomly sampled from Uniform$[0, 1]$ distribution. The three right censoring variables are also Exponentially distributed with three different chosen rates to create preset censoring rates. For CEs calculation, all covariates are categorized into 10 uniformed bins, and the response variable T is also categorized into 10 bins as well based on their Kaplan-Meier estimates. Each simulated data set is 10,000 ($= n$) in size.

[**Experiment-:** 10% **censoring rate**]We report our CE evaluations in Table 4.4. The row-wise CEs are ranked from the top-to-bottom in the three feature settings, respectively. In the 1-feature setting, we see only V1 and V7 having significant SCE-drops. So they are individual order-1 major factor candidates. The CEs of V2 and V3 are even as low as that of those random noise features.

The interacting effect of V2 and V3 are visible in the 2-feature setting by having a SCE-drop being many times of CEs of either V2 or V3. That is, feature-pair (V2, V3) is an order-2 major factor candidate. The SCE-drop of the feature-pair (V1, V7) is larger than V7's SCE-drop. This ecological effect indicates that V1 and V7 are currently present in the dynamics of response. So they are likely to be the order-1 major factors together. It is noted that CEs and SCE-drops of all the feature pairs of random noise provide the baseline of comparison. In the 3-feature setting, the feature triplets (V1, V2, V3) and (V2, V3, V7) achieve the lowest CEs and their SCE-drops indicate the ecological effects being achieved. That is, these three major factors: (V2, V3), V1, and V7, are concurrently present within the dynamics of T. Despite increasing uncertainty along with increasing censoring rates, exactly the same conclusions of major factors can be drawn from Table 4.5 and Table 4.6 based on the two simulated data sets with 20% and 30% censoring rates, respectively.

As for the Cox PH results, V1 and V7 are significant across the three experiments, while in contrast the two features: V2 and V3, are never seen as simultaneously significant in the three experiments. We only observe that V3 is somehow significant in experiment-1, V2 in experiment-2. But both features are insignificant in experiment-3. These experimental results converge to the fact that any interacting effect with non-product format is likely ignored by Cox PH results, especially when the censoring is high.

1-feature	CE	SCE-dp	p-value(PH)	2-feature	CE	SCE-dp	3-feature	CE	SCE-dp
V7	1.0103	0.0158	<2e-16	V2_V3	0.9292	0.0955	V1_V2_V3	0.8114	0.1178
V1	1.0156	0.0105	<2e-16	V1_V7	0.9898	0.0205	V2_V3_V7	0.812	0.1172
V9	1.0246	0.0014	0.7942	V2_V7	0.9973	0.013	V2_V3_V10	0.8217	0.1075
V3	1.0247	0.0013	0.0515	V7_V8	0.9982	0.0121	V2_V3_V5	0.8221	0.1072
V2	1.0247	0.0013	0.3575	V3_V7	0.9984	0.0118	V2_V3_V4	0.8238	0.1054
V4	1.0248	0.0013	0.5773	V7_V9	0.9993	0.011	V2_V3_V8	0.8267	0.1026
V10	1.0248	0.0012	0.7622	V5_V7	0.9994	0.0108	V2_V3_V9	0.8274	0.1018
V6	1.0249	0.0012	0.654	V6_V7	0.9996	0.0106	V2_V3_V6	0.8282	0.101
V8	1.0251	0.001	0.9028	V7_V10	0.9999	0.0104	V1_V7_V8	0.8728	0.117
V5	1.0251	0.001	0.4253	V4_V7	1.0004	0.0099	V1_V2_V7	0.8743	0.1155

Table 4.4. Experiment-1 (censoring rate is 10%): Ranked conditional entropies (CE) and successive CE-drop for selected feature-sets; "SCE-dp" short for "SCE-drop"; "p-value(PH)" for the fitted Cox PH model.

1-feature	CE	SCE-dp	p-value(PH)	2-feature	CE	SCE-dp	3-feature	CE	SCE-dp
V7	1.0759	0.0116	<2e-16	V2_V3	1.0034	0.0827	V2_V3_V7	0.8997	0.1037
V1	1.0773	0.0102	<2e-16	V1_V7	1.0563	0.0196	V1_V2_V3	0.9032	0.1002
V10	1.0861	0.0014	0.5769	V1_V10	1.0649	0.0124	V2_V3_V6	0.9056	0.0978
V3	1.0862	0.0013	0.4753	V1_V9	1.0652	0.0121	V2_V3_V10	0.9058	0.0977
V8	1.0866	9e-04	0.7523	V7_V8	1.0652	0.0107	V2_V3_V8	0.9108	0.0927
V2	1.0866	8e-04	0.0866	V3_V7	1.0652	0.0107	V2_V3_V4	0.9122	0.0913
V6	1.0867	8e-04	0.9617	V7_V10	1.0657	0.0102	V2_V3_V5	0.9136	0.0898
V9	1.0868	7e-04	0.5642	V4_V7	1.0658	0.0101	V2_V3_V9	0.9174	0.0861
V4	1.0868	7e-04	0.7306	V6_V7	1.0658	0.0101	V1_V7_V10	0.9545	0.1017
V5	1.0869	6e-04	0.986	V1_V5	1.0665	0.0108	V1_V7_V9	0.9553	0.101

Table 4.5. Experiment-2 (censoring rate is 20%): Ranked conditional entropies (CE) and successive CE-drop for selected feature-sets; "SCE-dp" short for "SCE-drop"; "p-value(PH)" for the fitted Cox PH model.

4.5. ADNI data analysis

For the CEDA paradigm, the scheme of data categorization for all quantitative variables is given as follows. Measurements of each quantitative covariate feature are grouped into 4 equal-spaced bins. As we only have 903 subjects in total, 4 bins would be an appropriate choice to conduct the CEDA analysis to avoid the effect of the curse of dimensionality. There are 2 binary categorical features, GENDER ($V2$) and FLDSTRENG-bl ($V5$). As for the T_i and C_i which have their ranges within $[6, 162]$ (month). It is noted that the maximum observed T_i is 138 and the

1-feature	CE	SCE-dp	p-value(PH)	2-feature	CE	SCE-dp	3-feature	CE	SCE-dp
V1	1.0996	0.0096	<2e-16	V2_V3	1.0466	0.0616	V1_V2_V3	0.9561	0.0906
V7	1.1007	0.0085	<2e-16	V1_V7	1.0851	0.0146	V2_V3_V7	0.9609	0.0858
V9	1.1078	0.0014	0.697	V1_V8	1.0894	0.0102	V2_V3_V8	0.9647	0.0819
V6	1.1081	0.0011	0.436	V1_V2	1.0899	0.0097	V2_V3_V6	0.965	0.0816
V10	1.1082	0.001	0.871	V1_V9	1.0904	0.0093	V2_V3_V5	0.9667	0.08
V2	1.1083	0.001	0.988	V1_V4	1.0906	0.009	V2_V3_V10	0.9676	0.079
V8	1.1084	8e-04	0.562	V7_V9	1.0908	0.0099	V2_V3_V4	0.9679	0.0788
V5	1.1084	8e-04	0.877	V1_V10	1.0908	0.0088	V2_V3_V9	0.9683	0.0784
V4	1.1087	6e-04	0.808	V1_V5	1.0912	0.0084	V1_V3_V7	0.9981	0.0869
V3	1.1089	3e-04	0.323	V6_V7	1.0912	0.0095	V1_V7_V8	0.9996	0.0854

Table 4.6. Experiment-3 (censoring rate is 30%): Ranked conditional entropies (CE) and successive CE-drop for selected feature-sets; "SCE-dp" short for "SCE-drop"; "p-value(PH)" for the fitted Cox PH model.

observed C_i is 162. We opt to obtain time bins by the following scheme: dividing $[6, 162]$ into 4 bins: $[6, 46), [46, 85), [85, 139), [139, 163)$ so that all observed T_i are included in the first three bins. In this way, we are able to clearly learn the structure and characteristics between the observed time (T_i) and censoring time (C_i).

Denote the right censoring ADNI data set as $\{(Y_i, \mathcal{V}_i, \delta_i)|i = 1, .., n\}$ with $Y_i = (T_i \wedge C_i)$, $\mathcal{V}_i = \{V1_i, .., V16_i\}$, δ_i the censoring status and sample size $n = 903$. If the ith subject is uncensored, $\delta_i = 1$, then $Y_i = T_i(< C_i)$ is its observed survival time, while if the ith subject is censored, $\delta_i = 0$, then $Y_i = C_i(< T_i)$ is the censoring time defined as the exam date of first no-show. The total number of uncensored data points is $n_o = \sum_{i=1}^{903} \delta_i = 346$, while the total number of censored data points is $n_c = n - n_o = 557$. So, this data set's censoring rate is over 61%.

Consider the two ensembles of observed censoring and survival times: $\{C_i|\delta_i = 0, i = 1, .., 557\}$ and $\{T_i|\delta_i = 1, i = 1, .., 346\}$. It is noted the largest C_i among the censored data point is $\max\{C_i|\delta_i = 0, i = 1, ..., 557\} = 162$, while the largest T_i among the uncensored data point is $\max\{T_i|\delta_i = 0, i = 1, ..., 346\} = 138$. Therefore, we have to take the convention that the largest censored data is taken as an uncensored one as usually done within many computational operations in Survival Analysis, for instance, in constructing Kaplan-Meier estimation of the survival function [**57**]. We make use of this convention because the redistribution-to-the-right algorithm developed by B. Efron in [**26**] is heavily applied in this paper. With respect to these two ensembles, we report their histograms with respect to four bins: $[6, 46)$, $[46, 85)$, $[85, 139)$, $[139, 163)$, in Figure 4.1.

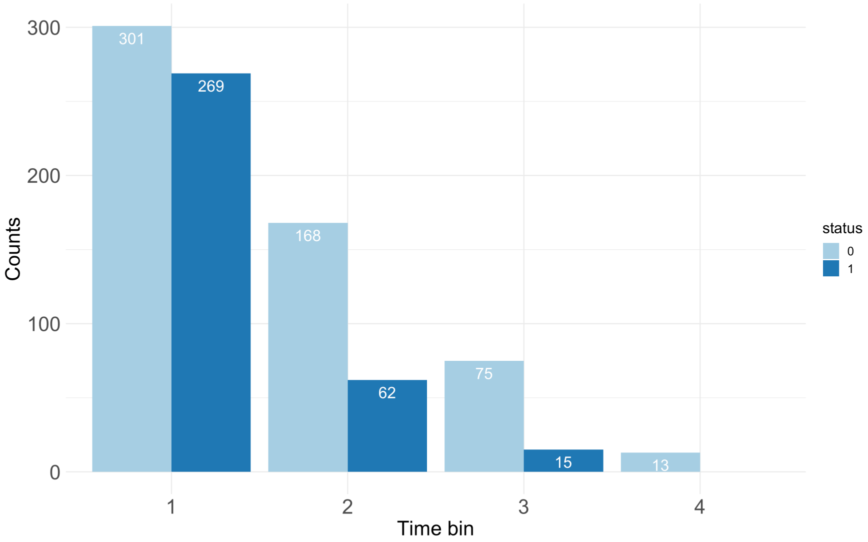


FIGURE 4.1. Histograms of 557 C_i (censoring time) (in dark blue color bars) and 346 observed T_i (in light-blue bars).

Since each individual subject's potential T_i and C_i are to be realized and observed at a common setting: a scheduled examination date, it becomes not at all obvious whether the T_i is indeed stochastically independent of C_i. Needed rigorous testing is constructed and conducted in the next subsection. The efforts invested in confirming this answer are essential because either positive or negative answers to this required fundamental assumption are expected to have significant impacts on the validity of any applications of the Cox Proportional Hazard model, which would be briefly reviewed below.

The Cox Proportional Hazard model, which was proposed by D.R. Cox in his 1972 landmark paper [**20**], is the most fundamental and the most popular modeling structure employed in the statistical topic of Survival Analysis. It is also widely used in analyzing data derived from ADNI as well. The classic version of the Cox proportional hazard (PH) model on the right censored data set is described as follows. Given $\{(Y_i, \mathcal{V}_i, \delta_i)|i = 1, .., n\}$ and assuming T_i being stochastically independent of C_i, the hazard rate function $\lambda_{\tilde{T}_i}(t)$ of $\tilde{T}_i$ is specified as:

$$\lambda_{T_i}(t) = \lambda_0(t)e^{\sum_{k=1}^{16} \beta_k V k_i},$$

for all $i = 1, .., n$. The assumed structural linearity and additivity are designed to accommodate all effects of the 16 covariate variables. The exponent of $\lambda_{T_i}(t)$ certainly can involve interacting effects among the 16 covariate features effects. Nevertheless, given no prior knowledge of which forms of interacting effects pertain to which pairs or triplets of features, the inclusion of interacting effects would result in an unrealistic and complex model.

Under the PH model structures and non-informative censoring assumption, the partial likelihood approach proposed in [**20**] is still the most widely used inference methodology in Survival Analysis. Nevertheless, it is worth reiterating that the global structure embraced by this PH model indeed is built upon an assumed homogeneity across the entire population contained in ADNI. From a rigorous standpoint, this homogeneous assumption is neither natural nor scientific. Though it might be practical, it is just parsimonious at best. Since this data set from ADNI is subject to two characteristics that could significantly impact results derived from Cox proportional hazard model structure. These two characteristics are: 1) the presence of heavy censoring; 2) the hidden heterogeneity among subjects. Both characteristics likely violate the validity regarding the global structures. In the last three subsections of this section, the partial likelihood results of the Cox proportional hazard model are compared with results derived based on our major factor selections on two scales: global and local.

4.5.1. Redistribute-to-the-right weight matrix. In this data analysis, all computations are primarily performed on the platform of the contingency table. This platform explicitly and correctly facilitates all calculations of measurements of conditional entropy, as such, we evaluate directed or indirect associations between the survival time T_i and any other covariate variables. When building a contingency table for such association involving variable T_i, we need to adopt a convention in Survival Analysis: the censored subject having its censoring time being equal to $\max\{C_j|\delta_j = 0, j = 1, ..., 557\} = 162$ is converted into an uncensored subject because this censoring time is beyond $\max\{T_i|\delta_i = 1, i = 1, ..., 346\} = 138$. With this convention, we illustrate how to build a 903×347 weight-matrix by applying the [Redistribution-to-the-right] algorithm discussed in Section 4.3 in this subsection. This weight matrix would the basis for building a contingency table pertaining to any covariate features or feature-sets against variable T_i.

For illustrative purposes, we build a 903×347 matrix lattice for distributed weights pertaining to the variable $\{Y_i\}$ against the variable $\{T_i\}$. By breaking all ties, we make all values of $\{Y_i|i =$

$1, .., 903\}$ distinct, and then ordered and arranged them from bottom-to-top along this matrix's row-axis, while all observed $\{T_{i'}|\delta_{i'} = 1, i' = 1, .., 347\}$ are made distinct, ordered and arranged from left-to-right along this matrix's column-axis. It is noted again here that the largest uncensored $T_{(347)} = 162$, which is originally censored.

Upon this 903×347 matrix lattice, each of its rows is constructed in the following fashion:

[$\delta_{(i)} = 1$**:** :] If the i-th ranked $Y_{(i)}$ is uncensored ($\delta_{(i)} = 1$), then the weight 1 goes to the i'-th column of $T_{(i')} = Y_{(i)}$, which is color-marked as a red-dot in Figure 4.2;

[$\delta_{(i)} = 0$**:** :] If the i-th ranked $Y_{(i)}$ is censored ($\delta_{(i)} = 0$), then the weight 1 would be re-distributed to all columns of $T_{(i')} > Y_{(i)}$ according to the [Redistribution-to-the right] algorithm, are represented by in changing-color-segment in Figure 4.2.

Let this matrix of distributed weights be denoted by $\mathcal{W}[Y, T]$, which is displayed in Figure 4.2.

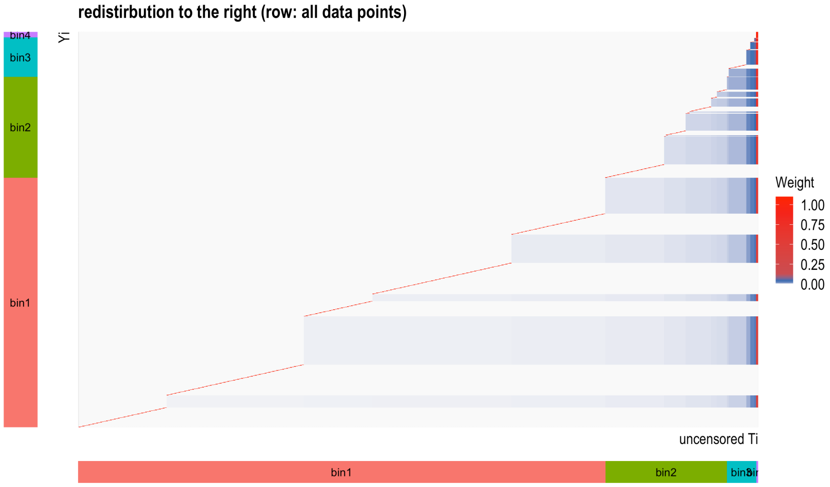


FIGURE 4.2. Matrix of distributed weights of individual 903 ranked Y_i against 347 (=346+1) observed survival time T_i; two color bars indicate the corresponding 4 divided time bins.

This weight matrix $\mathcal{W}[Y, T]$ would play a very fundamental role in all our applications of major factor selection protocol throughout this data analysis from here on. Since we can replace this feature of time variable Y_i with any one of the 16 features or feature-sets. That is, each row of $\mathcal{W}[Y, T]$ is identified via the subject's ID and its censoring status δ_i. As such, a weight matrix

$\mathcal{W}[Vk, T]$ is obtained by permuting the rows according to the increasing orders with respect to ordered values of Vk.

For any $k = 1, .., 16$, on the column-axis of $\mathcal{W}[Vk, T]$, the values $\{T_{(i')}|\delta_{(i')} = 1, i' = 1, .., 347\}$ are grouped into four bins: $[6, 46)$, $[46, 85)$, $[85, 139)$, $[139, 163)$. It is noted that the bin $[139, 163]$ contains only censored data points. Likewise, upon the row-axis of $\mathcal{W}[Vk, T]$, the values $\{V_{k_{(i)}}|i = 1, .., 903\}$ are also grouped into the four bins, which can be determined with respect to a histogram of Vk.

4.5.2. Testing non-informative censoring assumption. For testing the non-informative censoring assumption, we need to explore the associative relation between survival time (T_i) and censoring time (C_i). However, we only observed $Y_i == (T_i \wedge C_i)$ as the minimum T_i and C_i. That is, when $\delta_i = 0$, T_i is censored by C_i. But when $\delta_i = 1$, C_i is censored by T_i. The dual roles of T_i and C_i make this missing-data mechanism symmetric. Therefore, when $\delta_i = 0$ and observed C_i, we figure out where the missing bivariate(C_i, T_i) could potentially located by using the [Redistribution-to-the-right] algorithm. The 557×347 matrix of distributed weights, denoted as $\mathcal{W}[C, T]_{\delta_i=0}$, is reported in panel (a) of Figure 4.3. This weight matrix is indeed obtained simply by deleting the 346 rows of $\mathcal{W}[Y, T]$ corresponding to $\delta_i = 1$.

In contrast, when $\delta_i = 1$ and observed T_i, we again figure out where the missing bivariate(C_i, T_i) could potentially be located by using the [Redistribution-to-the-right] algorithm. The 346×557 matrix of distributed weights, denoted as $\mathcal{W}[T, C||_{\delta_i=1}$, is reported in panel (b) of Figure 4.3.

By applying the binning-scheme with respect to the four interval regions: $[6, 46)$, $[46, 85)$, $[85, 139)$, $[139, 163)$, on both axes of $\mathcal{W}[C, T]_{\delta_i=0}$ and transposed matrix $\mathcal{W}^T[T, C||_{\delta_i=1}$, we obtain two 4×4 contingency tables, as reported in 4x4tableCT. The contingency table $\mathcal{C}[C, T]$ denotes the sum of these two 4×4 contingency tables.

This contingency table $\mathcal{C}[C, T]$ reported in Table 4.7 indeed manifests the multiple aspects of associative relations between the censoring time C_i and survival time T_i. From the row- and column-wise aspects, we calculate the row-wise and column-wise Shannon conditional entropies and re-scale them by Shannon entropies of vectors of proportions of column-sums and row-sums, accordingly. Hypothetically, if a row-wise re-scaled CE is close to 1, then we know the information

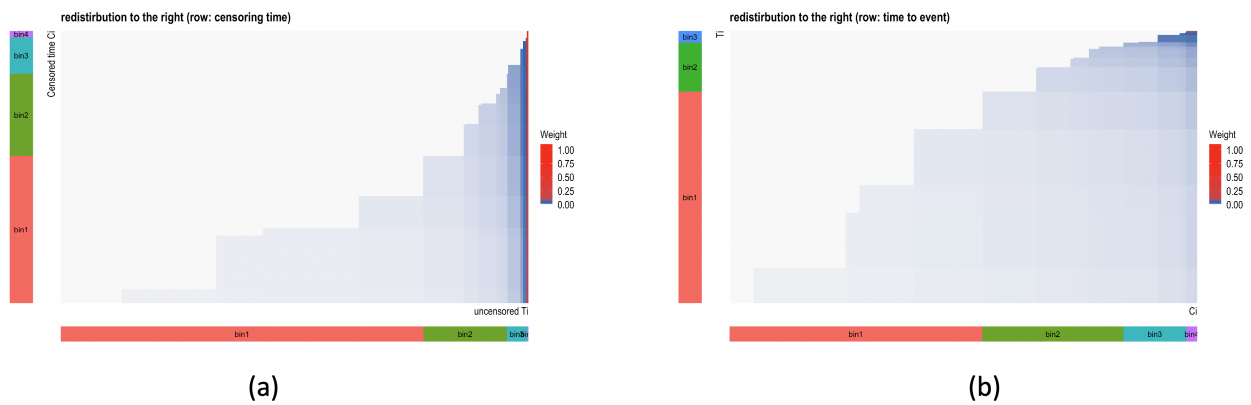


FIGURE 4.3. (a) Matrix of distributed weights of individual 557 ranked C_i against 347 (=346+1) observed survival time T_i and (b) matrix of distributed weights of individual 346 ranked T_i against 557 observed censored time C_i; color bars indicate the corresponding 4 divided time bins.

bin-T / bin-C	1	2	3	4
1	46.70	60.58	54.15	139.56
2	0.00	17.33	42.12	108.55
3	0.00	0.00	28.79	61.21
4	0.00	0.00	0.00	13.00
1	58.88	0.00	0.00	0.00
2	122.22	22.53	0.00	0.00
3	72.09	31.65	7.60	0.00
4	17.81	7.82	7.40	0.00

TABLE 4.7. Contingency table $\mathcal{C}[C,T]$: the sum of the 4×4 contingency table derived from 557×347 weight matrix $\mathcal{W}[C,T]_{\delta_i=0}$ (upper half) and the 4×4 contingency table derived from 346×557 weight matrix $\mathcal{W}^T[T,C|]_{\delta_i=1}$ (lower half).

about C_i not helping us in predicting T_i, while a column-wise re-scaled CE is away from 1 and smaller than 1, then we know the information of T_i indeed helping us in predicting C_i.

Our testing results from row-aspect are displayed in Figure 4.4 and testing results from column-aspect are displayed in Figure 4.5. Based on the Multinomial random mechanism, we simulate four row-wise alternative distributions against the null distribution based on the vector of column-sums as displayed in Figure 4.4. All five distributions heavily overlap. Therefore, the sum of Type-I and Type-II errors is large. Likewise, as displayed in Figure 4.5, similar results are found through the four column-wise alternative distributions against the null distribution based on the vector of row-sums. Thus, we conclude that associative relations between the censoring time C_i and survival time T_i are not evident. The non-informative censoring assumption stands.

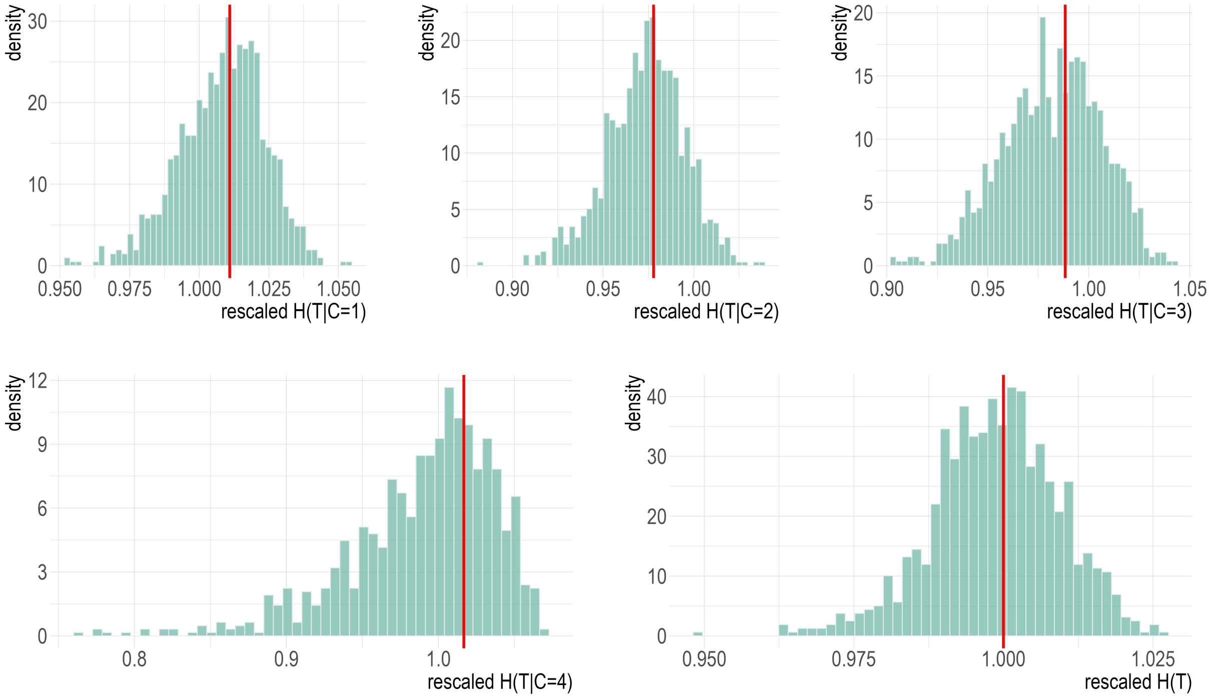


FIGURE 4.4. Four row-wise simulated alternative distributions marked with vertical lines at observed re-scaled CEs $\{1.0109,\ 0.9778,\ 0.9883,\ 1.0167\}$ against one simulated null distribution marked with one vertical line at 1.00. The entropy of vector of column-sum proportions is 1.2979.

The critical implication derived from the confirmation of the noninformative censoring scheme is regarding which response variables, either T_i or (T_i, C_i), are legitimate and should be used in Survival analysis. Since T_i is stochastically independent of C_i, we are not required to use (T_i, C_i) as the legitimate 2D response variable, or to use the categorical variable defined by the 16 categories in Table 4.7. That, we only need to use the categorized variable of T_i defined by the 4D vector of column-sums in Table 4.7 throughout the categorical exploratory data analysis (CEDA) carried out in this paper.

4.5.3. Using δ_i as a response variable. After confirming the fundamental assumption of non-informative censoring, we first look into the issue of whether the observed data of censoring

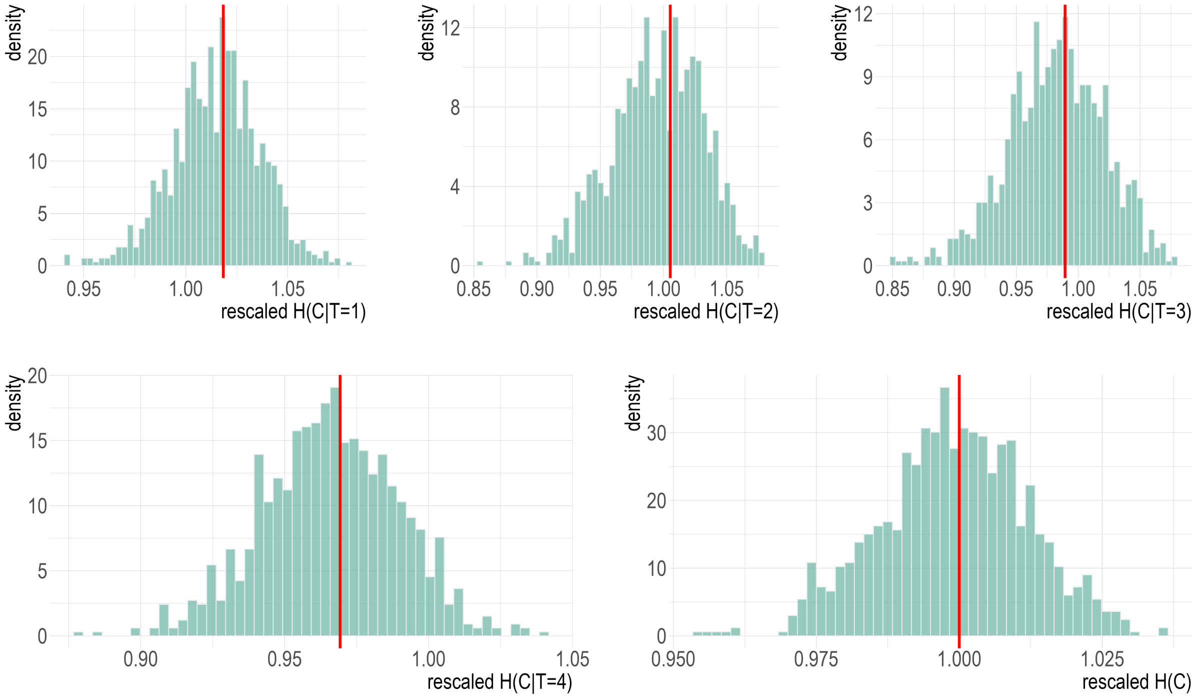


FIGURE 4.5. Four column-wise simulated alternative distributions marked with vertical lines at observed re-scaled CEs $\{1.0184, 1.0053, 0.9894, 0.9693\}$ against one simulated null distribution marked with one vertical line at 1.00. The entropy of vector of row-sum proportions is 1.2111.

status $\{\delta_i | i = 1, .., n\}$ could indeed shed light on the dynamics of T_i. The rationale is that $\delta_i = 1_{[T_i \leq C_i]} = 1 - 1_{[T_i > C_i]}$ is categorical transformation of T_i. Since T_i is stochastically independent of C_i, this transformed data logically should still carry relevant information. So, as the first step of our categorical exploratory data analysis (CEDA), we seek for any covariate variables $\{V1, .., V16\}$ that is highly associated with δ_i. We perform our CE computations by employing δ_i as a binary response variable and all covariate variables are categorized to have 4 bins based on their individual histograms. The relevant CEs are reported in Table 4.8.

From Table 4.8, we see that δ_i is apparently associated with V9 (MEM-mean) and V8 (ADAS13.bl), both of which achieve about 23% and 13% reductions of δ_i's uncertainty, respectively. This result implies that the directed association from T_i to the 16 covariate features surely would be much stronger than the results reported strongly.

1-feature	Feature name	CE	SCE-drop
V9	MEM-mean	0.5137	0.1518
V8	ADAS13.bl	0.5731	0.0924
V7	FAQ	0.6154	0.0501
V6	CDRSB.bl	0.6156	0.0500
V13	LAN-mean	0.6200	0.0455
V11	EXF-mean	0.6207	0.0448
V4	APOE4	0.6420	0.0236
V16	VSP-std	0.6530	0.0125
V15	VSP-mean	0.6560	0.0095
V5	FLDSTRENG.bl	0.6571	0.0084
V1	AGE	0.6575	0.0080
V12	EXF-std	0.6595	0.0060
V10	MEM-std	0.6600	0.0055
V14	LAN-std	0.6650	0.0006
V3	PTEDUCAT	0.6650	0.0005
V2	PTGENDER	0.6653	0.0002

TABLE 4.8. Conditional entropies (CE) of 16 features with censoring status δ_i as a response variable.

4.5.4. Computing major factors. Before exploring the directed association of all 16 covariate variables $\{V1, .., V16\}$ to response variable T_i, we display mutual association among these 16 covariate features via a 16 × 16 heatmap and a network of 16 nodes in the two panels of Figure 4.6. From panel (a), we see two highly associated feature-pairs: {V9, V8} and {V15, V16}, two moderately associated feature-pairs:{V6, V7}, {V11, V13}, and moderately associated feature-triplet:{V10, V12, V14}. These feature-sets are also mutually associated with varying degrees. In contrast, panel (b) reveals a global picture of the association among these 16 features with respect to a chosen threshold.

It is worth reiterating that, as mentioned in Section 4.4 and detailed in [**36**], any highly associative relationships among features would require extra computational efforts for explicitly clarifying and evaluating their joint conditional dependency or interacting effects much harder. We demonstrate why such required efforts are needed in this subsection by implementing our major factor selection protocol facilitated by contingency-table-based CE computations within this ADNI data set. Nonetheless, as would be discussed in the two subsections after this one, such efforts are indeed needed anyway for exploring heterogeneity within data.

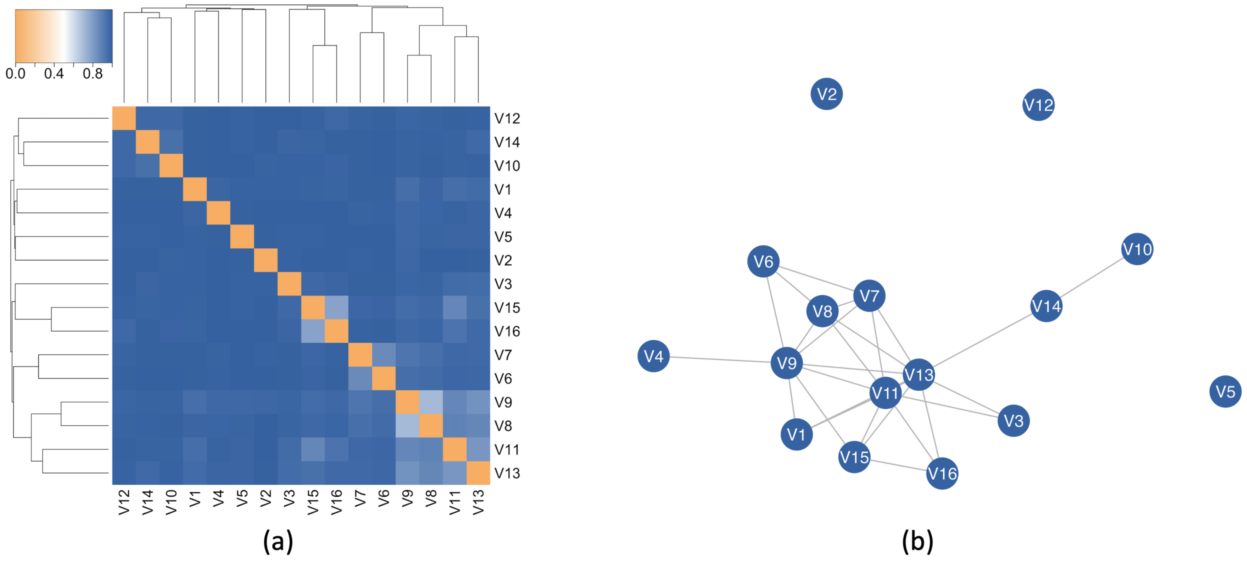


FIGURE 4.6. Heatmap (a) and network (b) of 16 covariate features in terms of Mutual Conditional Entropy (MCE). The threshold of the linkage is 0.97.

The first step of the major factor selection protocol is performed by calculating the conditional entropies (CEs) of all possible feature-sets of the collection of 16 covariate features from the 1-feature setting up to the 3-feature setting. Our CEs computations end at the 3-feature setting is a necessary choice for reflecting the uncensored sample size 346. All CEs of the 1-feature setting is reported in Table 4.9 coupled with results derived from Cox PH models on an individual feature basis. Specifically, we also report p-values of partial likelihood estimates of $\{\beta_1, .., \beta_{16}\}$ parameters. As for CEs of 2-feature and 3-feature settings, we only select and report 12 primary top-ranked pairs and triplets in Table 4.10 due to the large numbers of feature-pairs and feature-triplets.

Before comparing CEs and Cox's PH results, it is necessary to keep in mind that each CE is pertaining to one individual feature, while a p-value of covariate feature is evaluated under a global PH modeling structure. Therefore, these 16 parameter estimates are correlated, so are their p-values. From Table 4.9, we observe that feature: MEM-mean (V9), achieves the lowest CE, while ADAS13.bl (V8) is ranked 2nd. Both are potential order-1 major factor candidates, though they are not likely concurrently present when interpreting reduction of marginal uncertainty of T_i. That is, they could potentially join together in reducing uncertainty of T_i under conditional settings, as would be seen in the last two subsections regarding heterogeneity from both V9 and

1-feature	Feature name	CE	SCE-drop	p-value(PH)
V9	MEM-mean	1.1581	0.1397	<2e-16
V8	ADAS13.bl	1.1954	0.1023	0.5007
V11	EXF-mean	1.2354	0.0624	0.0007
V7	FAQ	1.2419	0.0559	<2e-4
V6	CDRSB.bl	1.2492	0.0486	<2e-4
V13	LAN-mean	1.2526	0.0452	0.5007
V4	APOE4	1.2697	0.0281	0.0013
V5	FLDSTRENG.bl	1.2849	0.0129	0.2386
V15	VSP-mean	1.2859	0.0118	0.2343
V14	LAN-std	1.2863	0.0115	0.0002
V12	EXF-std	1.2872	0.0106	0.0100
V10	MEM-std	1.2879	0.0099	0.7764
V1	AGE	1.2896	0.0082	0.6314
V16	VSP-std	1.2900	0.0078	0.0739
V3	PTEDUCAT	1.2952	0.0026	0.0025
V2	PTGENDER	1.2973	0.0005	0.1110

TABLE 4.9. Conditional entropies (CE) of 16 features with T as response variable, along with p-values of features via Cox Partial likelihood approach.

V8 perspectives. This possibility is in sharp contrast with PH results, which indicate V9 is more essential by having a p-value nearly zero than V8 with a p-value 0.5007. In other words, the act of dismissing the importance of V8 could come with a cost of ignoring important information of heterogeneity via V8 perspective, as would be seen in the last subsection of this section.

The EXF-mean(V11), FAQ (V7), CDRSB.bl (V6), LAN-mean (V13) and APOE4 (V4) are ranked from the 3rd to the 7th. Though these four features' individual CE-drops are significantly less than the CE-drops of the top two, their CE-drops are still significantly larger than CE-drops of the rest of the 10 features. Therefore, it is still possible for the individual features of { V6, V7, V11, V13, V4 } to be a candidate of stand-alone order-1 major factors. We also know the fact that {V6, V7} and {V11, V13} are moderately associated, while V4 is less associative with members of these two pairs. Thus, it is less likely that both V6 and V7 are concurrently present as two separate order-1 major factors, as are V11 and V13. On one hand, according to the PH results, V11 has a near zero p-value, but V13 has a p-value larger than 0.50. From the aspect of feature-pair {V11, V13}, the CEDA and PH results are coherent. On the other hand, both V6 and V7 have p-values near zeros. This PH result is not coherent with entropy-based CEDA results regarding the presence of order-1 major factors. The feature V4 has a p-value 0.0013 from the PH results. This computed PH result is not strongly incoherent from the CEDA perspective.

Behind these top 7 ranked features, the CE-drops of the rest of the 9 features are more or less falling into 2 tiers: {MEM-std (V10), VSP-mean (V15), FLDSTRENG.bl (V5), VSP-std (V16), EXF-std (V12), LAN-std (V14), AGE (V1)} for tier-1; {PTEDUCAT (V3), PTGENDER (V2)} for tier-2. From the CEDA perspective, feature members of these two tiers are increasingly less likely to be stand-alone order-1 major factors. But each one of them can couple with other features to become a potential candidate of order-2 major factor. In contrast, from the PH perspective, V14, V12, and V3 are highly significant by having p-values being equal to or less than 0.01. In Table 4.9, from the CEDA perspective, these features only achieve 0.9%, 0.8% and 0.2% of uncertainty reduction on T_i. Thus, the degrees of incoherence between PH and CEDA results are somehow evident.

Next, we turn the comparison to CEDA results of 2-feature or 3-feature settings in Table 4.10 with a focus on interacting effects among feature-pairs and feature-triplets. Some of the above conflicting results between CEDA-based major factor selection and Cox PH could be reconciled, but not all.

Before specifically discussing potential interacting effects in the 2-feature and 3-feature settings in Table 4.10, it is worth reiterating that an interacting effect of two or three features is referred to the confirmed presence of their conditional dependency given the response variable Y. Based on a relatively loose criterion in terms of SCE-drops, we identify and see how much extra uncertainty of response variable T_i is reduced by including a less potential feature variable relative to 3 or more times this feature's individual CE-drop, see details in [**14**,**36**]. A more strict criterion will be based on a criterion constructed based on the dominant feature. This strict version of the interacting effect is not used here. Further, the explicit form of interacting effect is left unknown when applying this criterion, it is somehow visible through the contingency table against the response variable T. In comparison, the task of confirming any conditional dependency of multiple orders is not at all simple under an assumed global structural model, such as Cox's PH model. In this paper, we explicitly demonstrate how to resolve this task.

Though the feature-set {V7, V9} achieves the lowest CE under the 2-feature setting, this pair's SCE-drop (0.0294) is less than the CE-drop (0.0559) of V7. This fact indicates that the conditional mutual information $I[V7, V9|T]$ is less than the marginal mutual information $I[V7, V9]$. That is, we can't confirm whether the feature-set {V7, V9} gives rise to a significant interacting effect or not. To confirm or dispute this fact. we either explicitly evaluate the mutual information $I[V7, V9]$, or evaluate how much extra information can V7 provide going beyond V9. We take the latter approach in the next subsection. Since we face the same issue for feature-sets {V6, V9}, {V9, V11}, {V8, V9} and {V4, V9}, which are the top-ranked 5 feature-pairs.

As for feature-set {V9, V10}, we know that V9 and V10 are marginally independent, so their marginal mutual information $I[V9, V10] = 0$. And the SCE-drop (0.0148) of V10 is larger than its CE-drop (0.0099) under the 1-feature setting. Therefore, their conditional mutual information $I[V9, V10|Y] \approx 0.0049$ is barely positive. Thus, according to the above criterion for confirming the interacting effect, the feature-set {V9, V10} has a very slight chance of being conditionally dependent given Y. However, the V10 could still play the role of assisting the order-1 major factor V9 in facilitating the information content of this ADNI data set or any predictive decision-making locally, as discussed in the next subsection.

In contrast, the three feature-sets: {V2, V9}, {V2, V8} and {V3, V8} have their SCE-drops being almost three times of V2 and V3's individual CE-drops or more. All involving CE-drops are confirmed to pass the reliability check. Therefore, we confirm the interacting effects of V2 and V3 with V8 at least. In fact, V1, V2 and V3 apparently have interacting effects with many features: V6, V7, V10, V11, V12, V15 and V16 as well. It is worth mentioning that some of the aforementioned interacting effects are indeed clinically confirmed and reported in [**4, 5, 25**].

In the 3-feature setting, we see that V9 is present in all the top-ranked feature-triplets. Its most often companion feature is either V7 or V11. There is a range of features for the third member of the triplet. Because of the marginal dependency of V9 on other features, the reliable pattern information becomes harder to be confirmed or disputed.

This phenomenon would be resolved to a great extent by subdividing the entire data collection with respect to categories of V9, which is called "de-associating" in [**36**]. As reported in the four tables given in the next subsection in the next subsection, by applying this simple computational procedure, we can take away or significantly reduce all covariate features' associations with V9. Overall, the rest of the 15 covariate features become less associative within each sub-collection of

2-feature	CE	SCE-drop	3-feature	CE	SCE-drop
V7_V9 (1)	1.1287	0.0294	V7_V9_V11 (1)	1.1002	0.0285
V6_V9 (2)	1.1346	0.0234	V6_V9_V11 (2)	1.1035	0.0311
V9_V11 (3)	1.1385	0.0196	V4_V7_V9 (3)	1.1039	0.0247
V8_V9(4)	1.1405	0.0175	V7_V8_V9 (4)	1.1076	0.0210
V4_V9 (5)	1.1418	0.0162	V7_V9_V12 (5)	1.1078	0.0209
V9_V10 (6)	1.1432	0.0148	V6_V7_V9 (6)	1.1089	0.0197
V1_V9(11)	1.1519	0.0061	V6_V9_V10 (7)	1.1090	0.0255
V3_V9(14)	1.1533	0.0047	V7_V9_V10 (8)	1.1095	0.0191
V2_V9(15)	1.1558	0.0022	V9_V11_V12 (9)	1.1099	0.0286
V8_V16(22)	1.1832	0.0121	V4_V9_V11 (10)	1.1108	0.0239
V3_V8(26)	1.1847	0.0106	V3_V7_V9 (16)	1.1129	0.0158
V2_V8(30)	1.1925	0.0029	V1_V6_V9 (20)	1.1144	0.0202

TABLE 4.10. Ranked conditional entropies (CE) and successive CE-drop for selected feature-sets.

study subjects defined by the 4 categories of V9, respectively. Another functional merit of de-associating is that this computing procedure allows us to figure out which feature variables can provide extra information beyond V9 while holding V9 constant. Therefore, this procedure is an effective way of discovering heterogeneity from the perspective of V9. We do the same heterogeneity exploration from the V8 perspective, as well as in the last subsection.

4.5.5. Heterogeneity w.r.t V9 (MEM-mean). {V9=1}:$199(n_o) - vs - 67(n_c)$**.** Upon the sub-collection: {V9=1}, of 266 subjects having the lowest ordinal category of MEM-mean (V9=1), we see the drastic ranking changes among the 14 covariate features in Table 4.11. The highest-ranked feature is V7, which is ranked 4th on the overall CE-drop in Table 4.9. The 2nd ranked one is V6, which is ranked 4th in Table 4.9, while the originally ranked 11th V16 is now ranked 4th. The V8 is ranked 3rd here, while it was ranked 2nd in Table 4.9. However, significant ranking drops are seen on features: V13, V4, and V5, which were previously ranked 6th to 8th. Now they are ranked 10th, 13th, and 15th here.

For reliability check on CE calculations, we generate a random $U[0,1]$ noise feature, denoted as V0, repeat this 200 times and calculate CEs: $H[Y|V0]$. The simulated null distribution for the 1-feature setting is given in the left of Figure 4.7. The top 6 ranked 1-features all have very small p-values. That is, their CEs are significant and real. The V1 has the p-value on the borderline. As for the 2-feature setting, as shown in the right of Figure 4.7, we found that only the top 3 feature-pairs have only borderline p-values. Nonetheless, we still carry out the task of identifying potential interacting effects as follows.

The most significant result observed from CEs of 1-feature and 2-feature is that majority of feature-pairs achieve the ecological effect, except the feature-pair {V6, V7}. This is a strong indication that all involving features are much less associated due to the de-associating procedure. Further, as listed below, we see many feature-pairs just like these three pairs: {V7, V15}, {V7, V13}, {V7, V3} and {V12, V14}, achieve SCE-drops that are at least three times of V15, V13, and V3's individual CE-drops. These results collectively suggest the following collection of 16 candidates of order-2 major factors:

$$\{\{V7, V3\}, \{V7, V4\}, \{V7, V13\}, \{V7, V15\}, \{V8, V3\},$$

$$\{V8, V4\}, \{V11, V4\}, \{V16, V2\}, \{V14, V12\}\{V1, V4\},$$

$$\{V10, V4\}, \{V12, V3\}, \{V12, V4\}, \{V12, V13\}, \{V2, V4\}, \{V13, V3\}\}.$$

Such results strongly indicate that not only members of {V7, V8, V11, V16, V1, V10} are concurrent candidates of order-1 major factor, but also they are coupled with members of {V4, V3, V2, V13, V15} to be candidates of order-2 major factors. That is, these feature-pairs indeed provide extra information beyond what V9 can provide at least in this sub-collection. This fact is the strongest evidence of heterogeneity that goes far beyond 2-feature setting of Table 4.11.

Next, it is also worth mentioning one sharp contrasting result: the feature-pair {V7, V8} achieves the ecological effect. So, V8 indeed can be concurrently present with V7 in reducing the uncertainty of T_i under the sub-collection {V9=1}. This result is indeed rather significant because V8 is an important feature in AD literature. It is also interesting that the feature V16 (VSP-std), which doesn't demonstrate a significant role in the overall setting at all in Table 4.9 and Table 4.10, surprisingly joins V7 and V8 in reducing the uncertainty of the response variable Y within this sub-collection of {V9=1}. This is the second piece of evidence of heterogeneity embedded within ADNI data.

The third piece of evidence of heterogeneity is collectively provided by the ranking changes pertaining to the originally lowest-ranked three features in Table 4.9: V1(Age), V3(PTEDUCAT), and V2(PTGENDER). They are now ranked 6th, 11th, and 12th, while V4's ranking changed from 7th to 10th.

When comparing CEDA and PH results, we clearly notice a few conflicting results observed from PH results reported in the 3rd column of Table 4.11. Given that feature-pair {V7, V6} doesn't achieve the ecological effect, their concurrent presence is not confirmed in CEDA analysis. However, PH results strongly indicate both features are significant simultaneously. While {V7, V8} achieves the ecological effect in CEDA analysis, on the contrary, the PH results indicate V8 is insignificant by having a p-value 0.1209. Further, results based on the PH model indicate features: {V11, V16, V1}, have p-values less than 0.05 and V14 has a p-value between 0.05 and 0.1, while the rest of the features are not significant. That is, PH results have completely missed all potential interacting effects like the 16 feature-pairs identified by our major factor selection protocol. This is a chief difference between CEDA and PH results.

1-feature	CE	SCE-drop	p-value(PH)	2-feature	CE	SCE-drop
V7	0.6112	0.0432	0.0001	V7_V11	0.5877	0.0235
V6	0.6307	0.0238	0.0461	V7_V16	0.5879	0.0233
V8	0.6347	0.0198	0.1209	V7_V8	0.5886	0.0226
V11	0.6364	0.0182	0.0205	V1_V7	0.5906	0.0206
V16	0.6375	0.0170	0.0004	V4_V7	0.5919	0.0193
V14	0.6390	0.0156	0.0714	V7_V14	0.5925	0.0187
V1	0.6442	0.0104	0.0002	V6_V7	0.5930	0.0182
V10	0.6461	0.0084	0.6311	V3_V7	0.5952	0.0160
V12	0.6502	0.0043	0.3741	V7_V10	0.5967	0.0145
V4	0.6503	0.0042	0.5398	V7_V15	0.5971	0.0141
V3	0.6506	0.0039	0.0334	V7_V12	0.5984	0.0128
V2	0.6506	0.0039	0.1480	V2_V7	0.5994	0.0118
V13	0.6511	0.0034	0.3634	V11_V14	0.6005	0.0358
V15	0.6534	0.0011	0.3666	V7_V13	0.6008	0.0104
V5	0.6542	0.0003	0.9179	V12_V14 (18)	0.6056	0.0333

TABLE 4.11. {V9=1}:Top 15 ranked conditional entropies (CE) and successive CE-drop under 1-feature and 2-feature settings. n = 266 with 199 uncensored data points.

What are the potential consequences that could be derived from the identified collection of 16 candidates of order-2 major factors? How these consequences of many faced heterogeneity embedded within data would impact our understanding of the dynamics of progression from MCI to AD? The full discussions of these two essential questions would be deferred to the ongoing

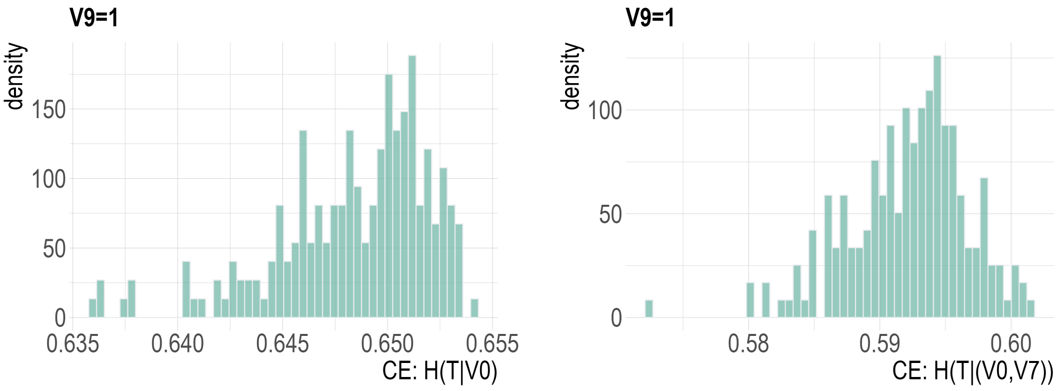
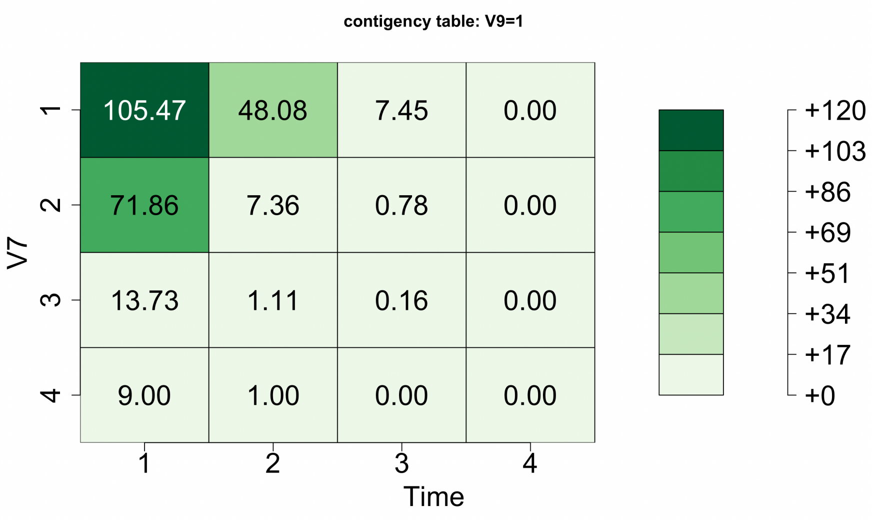


FIGURE 4.7. Null distributions for reliability checks for 1-feature and 2-feature settings: Left: simulated CEs of $H[T|V0]$; Right: simulated CEs of $H[T|(V0,V7)]$ in sub-collection {V9=1} based on 200 simulated $U[0,1]$ based features.

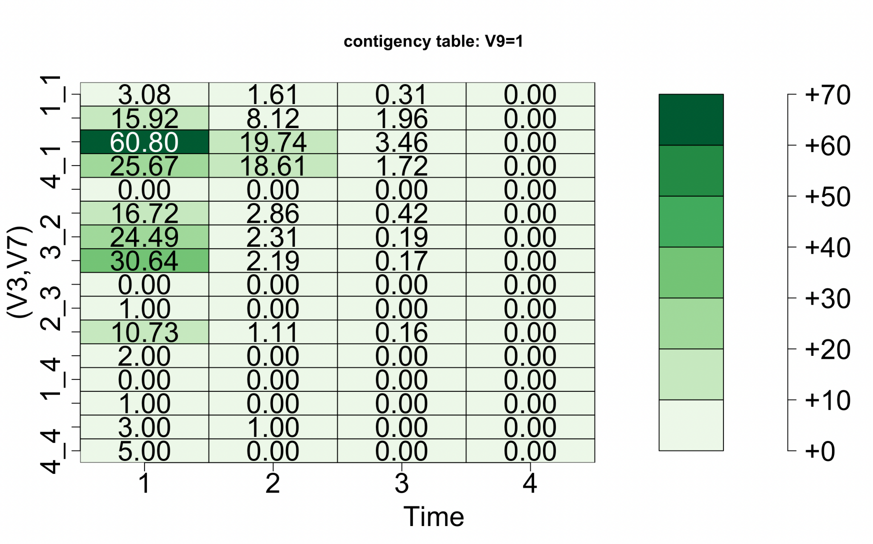
Part-II of this study. Here, we only briefly mention some clues leading toward the to-be-proposed resolutions.



V7 \ Time	1	2	3	4
1	105.47	48.08	7.45	0.00
2	71.86	7.36	0.78	0.00
3	13.73	1.11	0.16	0.00
4	9.00	1.00	0.00	0.00

FIGURE 4.8. Contingency table of $C[V7 - vs - T]$ within sub-collection $V9 = 1$.

Consider the contingency table $C[V7 - vs - T|V9 = 1]$ shown in Figure 4.8 as a way of precisely revealing the first layer of heterogeneity with respect to V7 within the sub-collection {V9=1}. We see relatively distinct 4 rows. In particular on the 2nd, 3rd and 4th rows, we intuitively predict T_i falling the category-1 of T if V7 is in categories {2, 3, 4}. This predictive decision-making is correct with varying correct rates and surely subject to some varying error rates across these three categories. In contrast, when V7 is seen in category {1}, then the correct rate goes down and the error rate goes up. Would the interacting effect of {V7, V3} help?



(V3,V7)	1	2	3	4
1_1	3.08	1.61	0.31	0.00
	15.92	8.12	1.96	0.00
	60.80	19.74	3.46	0.00
4_1	25.67	18.61	1.72	0.00
	0.00	0.00	0.00	0.00
	16.72	2.86	0.42	0.00
3_2	24.49	2.31	0.19	0.00
	30.64	2.19	0.17	0.00
	0.00	0.00	0.00	0.00
2_3	1.00	0.00	0.00	0.00
	10.73	1.11	0.16	0.00
	2.00	0.00	0.00	0.00
1_4	0.00	0.00	0.00	0.00
	1.00	0.00	0.00	0.00
	3.00	1.00	0.00	0.00
4_4	5.00	0.00	0.00	0.00

FIGURE 4.9. Contingency table of $C[(V3, V7) - vs - T]$ within the sub-collection $V9 = 1$.

The contingency table $C[(V3, V7) - vs - T|V9 = 1]$ is shown in Figure 4.9. Each category of V7 is divided into 3 or 4 bivariate-categories with respect to the four categories of V3. To precisely see which rows of V7 are being improved by bivariate-categories, we present the re-scaled CEs with respect to marginal CE of T within sub-collection {V9=1} in Figure 4.11: re-scaled CEs pertaining to categories of V7 (in blue dots), re-scaled CEs pertaining to categories of (V3,V7) (in orange dots). It is noted that there are overlapping dots at zeros in bivariate-categories: (2, 3), (4, 3), (2, 4), and (4,4). We propose to encode those subjects falling into these bivariate categories as, for example, V9-1-V7-3-V3-2-T1, V9-1-V7-3-V3-4-T1, V9-1-V7-4-V3-2-T1, and V9-1-V7-4-V3-4-T1. Such new code-ID indicates when a subject can be identified without the uncertainty of categories of T. If we relax the coding criterion just slightly to allow a small positive CE, such as 0.05, then we will also include subjects falling into bivariate-categories (3,3). The 12 subjects will be encoded with code-ID: V9-1-V7-3-V3-3-T1.

Further, even though feature-pair {V7, V6} doesn't achieve the ecological effects and just ranked above feature-pair {V7, V3}, we still can see many of their bivariate-categories achieving rather low or even zero CEs in Figure 4.10. Subjects falling into those bivariate-categories are qualified for short code-IDs. See also Figure 4.11, where are many overlapping dots at or very near zero CEs.

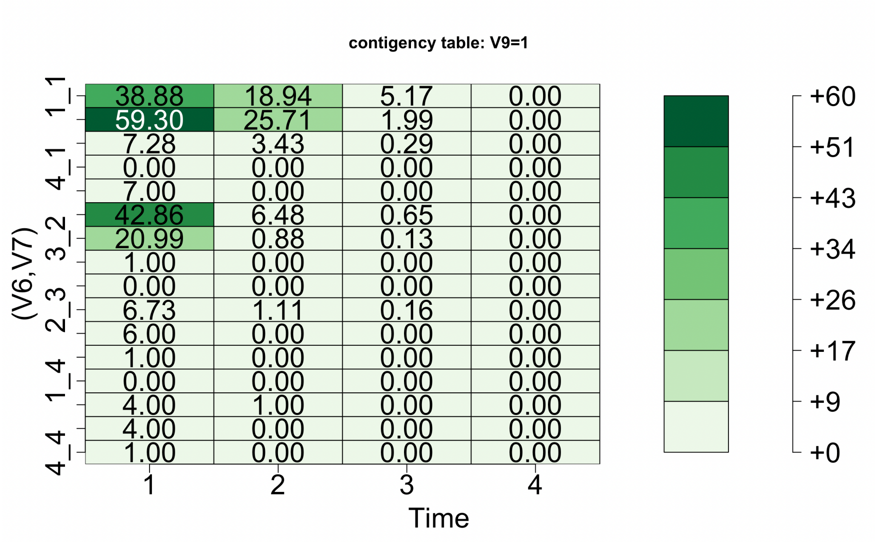


FIGURE 4.10. Contingency table of $C[(V6, V7) - vs - T]$ within the sub-collection $V9 = 1$.

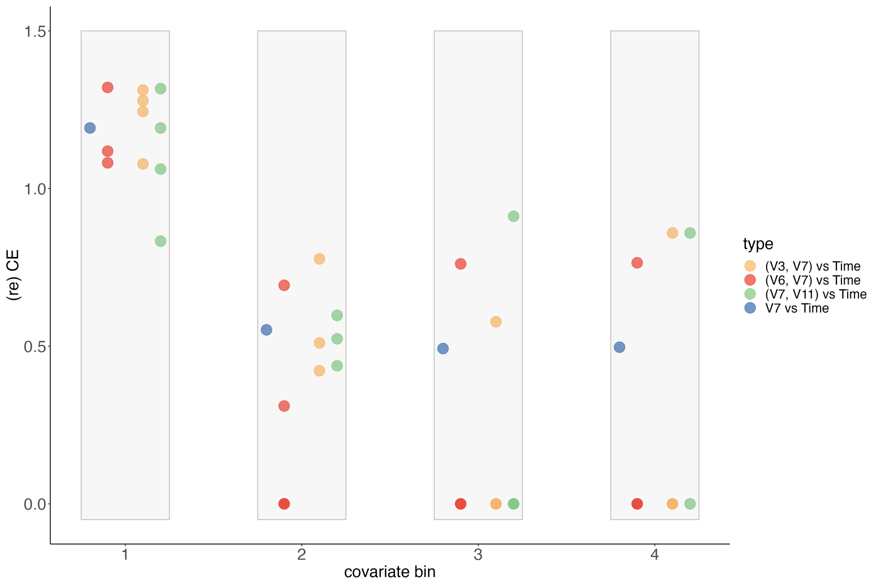


FIGURE 4.11. Heterogeneity of CE-expansions ($V9 = 1$): V7-vs-T (blue dots); (V7, V6)-vs-T (red dots); (V7, V3)-vs-T (orange dots); (V7, V11)-vs-T (green dots). The CE of T in sub-collection {V9=1} is equal to 0.6545. There are overlapping dots at and near zero.

{V9=2}:$141(n_o) - vs - 332(n_c)$**.** The sub-collection {V9=2} involves a heavy censoring rate. It is more than $\frac{2}{3}$. Our CE computations reported in Table 4.12 reveal very distinct ranking among 1-feature and 2-feature settings from that found on the sub-collection: {V9=1}. This distinction is an indication of a large-scale heterogeneity.

In the 1-feature setting, the lowest CE is surprisingly achieved by the feature: V4, and the 2nd and 3rd-ranked CEs are achieved by V6 and V7, respectively. Unlike in {V9=1}, V10 (MEM-std) rises to the 4th, which is ranked ahead of V11 and V8. V8 is ranked 3rd in in {V9=1}. We also had significant ranking drops for V16 and V1: 5th to 9th for V16; 7th to 14th for V1. These ranking drops would significantly impact the CEs of feature-pairs as seen in the 2-feature setting.

Also, for reliability check on CE calculations, we simulate a random $U[0,1]$ noise feature, denoted as V0, repeat this 200 times and calculate CEs: $H[Y|V0]$ as shown in the left of Figure 4.12. We see that, due to having more uncensored data points in this sub-collection {V9=2}, many more single features have very small p-values. As for the 2-feature setting, the middle and right of Figure 4.12 show the null distributions of simulated CEs of $H[T|(V0, V4)]$ and $H[T|(V0, V6)]$, respectively. Via panel (a), even the 14th ranked feature-pair (V4, V8) has a relatively smaller p-value, while via panel (b), the 15 ranked feature-pair (V7, V10) has a very small p-value. Again, we accordingly carry out the task of identifying potential interacting effects as follows.

For the 2-feature setting, the most striking pattern is that all $\binom{5}{2}(=10)$ feature-pairs of the top 5 ranked features: V4, V6, V7, V10, and V11, appear in the top 15 list and achieve the ecological effects. This computational fact indicates that the two members of each pair can be concurrently present as order-1 major factors. Such a phenomenon is made possible by the de-associating procedure, which makes the features less associated or even independent by taking off their association with V9. Further, the 17th ranked pair: {V3, V11}, demonstrates a strong interacting effect by having the SCE-drop of adding V3 being more than 5 times V3's CE-drop. Similar interacting effects with less strength are also seen, such as {V1, V6}, {V7, V13}, just to list a few.

These aforementioned order-1 features and order-2 feature-pairs are potential candidates for building their contingency tables against T in order to map out the heterogeneity within this sub-collection {V9=2}. However, this mapping out would not yield very decisive or precise predictions

1-feature	CE	SCE-drop	p-value(PH)	2-feature	CE	SCE-drop
V4	1.2897	0.0287	0.0000	V6_V11	1.2582	0.0348
V6	1.2931	0.0253	0.0002	V4_V7	1.2589	0.0308
V7	1.2948	0.0236	0.0000	V4_V12	1.2591	0.0306
V11	1.2981	0.0203	0.0842	V4_V6	1.2595	0.0302
V10	1.2986	0.0198	0.4121	V7_V11	1.2609	0.0339
V8	1.2998	0.0186	0.0102	V6_V10	1.2621	0.0309
V12	1.3058	0.0126	0.0885	V4_V11	1.2635	0.0262
V15	1.3086	0.0099	0.0298	V4_V10	1.2671	0.0226
V16	1.3128	0.0056	0.1459	V10_V11	1.2672	0.0310
V5	1.3129	0.0056	0.8022	V7_V12	1.2676	0.0272
V14	1.3129	0.0056	0.0713	V6_V15	1.2700	0.0230
V3	1.3144	0.0040	0.0014	V11_V12	1.2704	0.0277
V13	1.3146	0.0039	0.5264	V4_V8	1.2712	0.0185
V1	1.3163	0.0022	0.4214	V7_V10	1.2729	0.0219
V2	1.3117	0.0013	0.9507	V3_V11 (17)	1.2765	0.0216

TABLE 4.12. {V9=2}:Top 15 ranked conditional entropies (CE) and successive CE-drop under 1-feature and 2-feature settings. $n = 473$ with 141 uncensored data points.

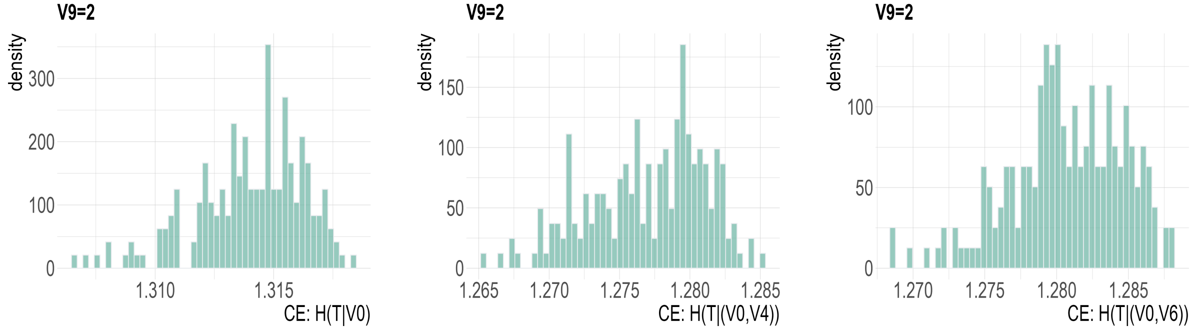


FIGURE 4.12. Null distributions for reliability checks at 2-feature settings: Left: simulated CEs of $H[T|(V0, V4)]$; Middle: simulated CEs of $H[T|(V0, V6)]$; Right: simulated CEs of $H[T|(V0, V6)]$, in sub-collection {V9=2} based 200 simulated $U[0, 1]$ based features.

because these features and feature-sets can't achieve significant uncertainty reductions. For instance, even the top-ranked feature-pair {V6, V11} can only achieve less than 5% reduction of uncertainty of T, which is calculated as $\frac{0.0348+0.0253}{0.0348+0.0253+1.2582}$. Details of this overall conclusion in this sub-collection {V9=2} can be seen through three figures: Figure 4.13 for contingency table $C[(V6, V11)-vs-T]$; Figure 4.14 for contingency table of $C[(V4, V12)-vs-T]$ and Figure 4.15 for

CE-expansion plots pertaining to contingency tables $C[(V6, V11)-vs-T]$ and $C[(V4, V12)-vs-T]$. Based on these three figures, we see only a few bivariate-categories can receive short code-IDs. We can make the same conclusion through the plots of CE-expansions pertaining to feature-pairs {V6, V11} and {V4, V12} presented in Figure 4.15.

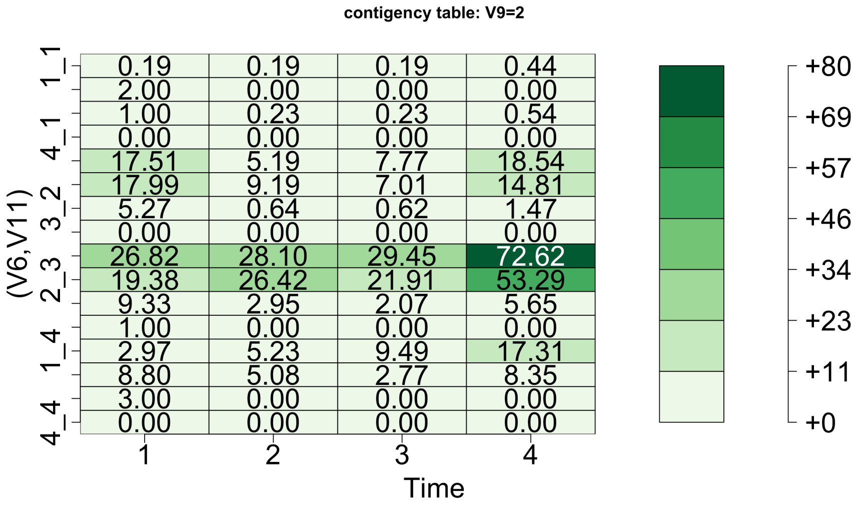


FIGURE 4.13. Contingency table of $C[(V6, V11) - vs - T]$ within the sub-collection $V9 = 2$.

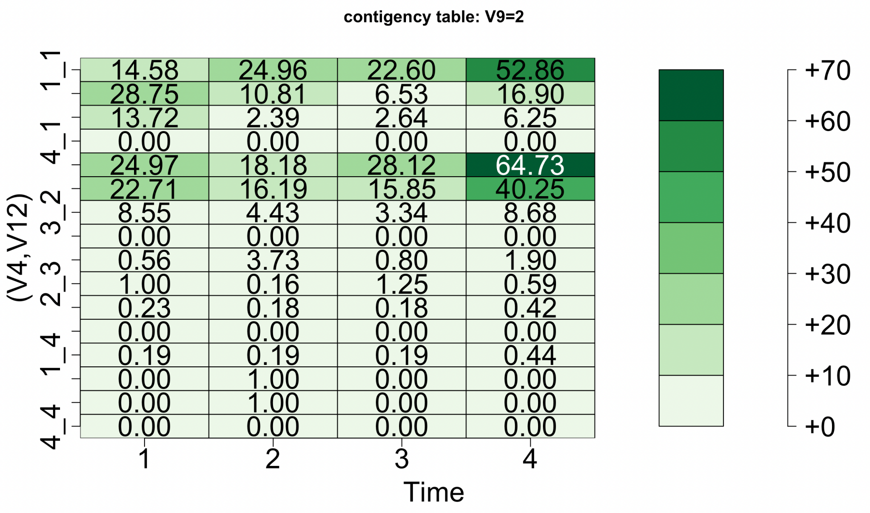


FIGURE 4.14. Contingency table of $C[(V4, V12) - vs - T]$ within the sub-collection $V9 = 2$.

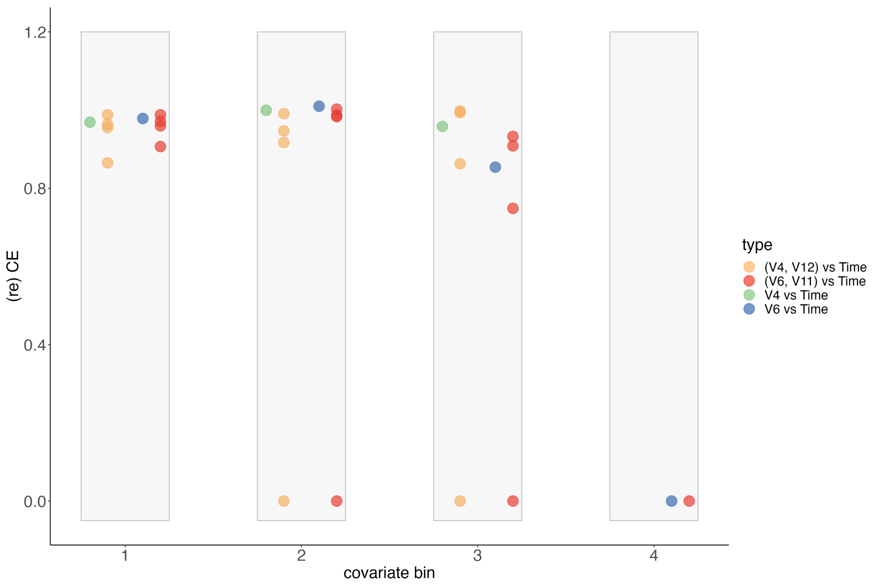


FIGURE 4.15. Heterogeneity of CE-expansion ($V9 = 2$) of V6-vs-T (blue dots) and V4-vs-T (green dots) and and (V6, V11)-vs-T (red dots) and (V4, V12)-vs-T (orange dots). The CE of T in sub-collection {V9=2} is equal to 1.3183.

Therefore, we can conclude that subjects in this sub-collection {V9=2} need different perspectives to look through them and analyze their data. For instance, we need to use different features or feature-sets to define sub-collections. We perform a different set of sub-collections defined by V8 in the next subsection.

For comparing the above CEDA results with the PH results, which are reported in the third column of Table 4.12, we see incoherent results from the aspects of {V10, V12, V15, V14, V3}. V10 is ranked 5th in CEDA results, but it is insignificant in PH results. While features: {V12, V15, V14, V3} are ranked 7th, 8th, 11th, and 12th, respectively, they are all significant in PH results. Further, PH results indicate that the 4th ranked V11 has a p-value much larger than the 12th ranked V3's, even though feature-pair {V3, V11} seems to have an interacting effect according to CEDA results. Nonetheless, the uncertainty reduction achieved by {V3, V11} is rather low. By summarizing these incoherent comparisons between CEDA and PH results, we suspect that PH results could be rather unreliable because of the presence of heavy censoring in this sub-collection {V9=2}. We further see its unreliability turning into incapability in the next sub-collections: {V9=3} and {V9=4}.

{V9=3}: $4(n_o) - vs - 147(n_c)$**.** In the sub-collection {V9=3}, there are only 4 uncensored and 147 censored data points. The PH hazard regression model becomes completely incapable of extracting any reliable information from data pertaining to this sub-collection. In sharp contrast, our contingency-table-based CEDA computations are not affected. Based on Table 4.13, the CE of T is about one-third of CE of T in sub-collection {V9=2} and one-half of CE of T in sub-collection {V9=1}. Further, some of the rest of the 15 features still offer reasonable amounts of information beyond the information of {V9=3}. This is one striking aspect of heterogeneity with respect to V9 and CEDA computations.

Based on Table 4.13, the top two ranked features are V7 and V6, respectively. So, the presences of V7 and V6 are among the top three ranked features across sub-collections: from {V9=1} to {V9=3}. This result strongly indicates that either V7 or V6 could provide extra information beyond V9 at least within these three sub-collections. The evidence of V7 is seen in Figure 4.16 having two rows with relatively low CEs: the 3rd and 4th. Furthermore, based on results of 1-feature setting in Table 4.13, the fact that V10 is ranked 4th is unseen in the sub-collections:{V9=1} and {V9=2}.

1-feature	CE	SCE-drop	p-value(PH)	2-feature	CE	SCE-drop
V7	0.4951	0.0429	0.0000	V7_V11	0.4699	0.0252
V6	0.5081	0.0299	0.0000	V6_V10	0.4734	0.0347
V11	0.5155	0.0225	0.0000	V7_V10	0.4743	0.0208
V10	0.5191	0.0188	0.0000	V7_V15	0.4781	0.0170
V4	0.5196	0.0183	0.9980	V1_V7	0.4801	0.0150
V8	0.5228	0.0152	0.0000	V6_V7	0.4829	0.0122
V1	0.5241	0.0139	0.0000	V4_V7	0.4841	0.0110
V3	0.5260	0.0120	0.0000	V3_V7	0.4856	0.0095
V15	0.5293	0.0087	0.0000	V6_V11	0.4864	0.0216
V5	0.5308	0.0072	0.0000	V7_V8	0.4865	0.0086
V13	0.5324	0.0056	0.0000	V7_V12	0.4871	0.0080
V14	0.5339	0.0041	0.8120	V10_V11	0.4874	0.0280
V16	0.5358	0.0022	0.0000	V2_V7	0.4875	0.0076
V12	0.5370	0.0010	0.0000	V7_V13	0.4887	0.0064
V2	0.5378	0.0002	0.0000	V4_V6 (17)	0.4898	0.0182

TABLE 4.13. {V9=3}:Top 15 ranked conditional entropies (CE) and successive CE-drop under 1-feature and 2-feature settings. $n = 147$ with 5 uncensored data points.

For the reliability check on CE calculations, we use the same simulation plan and calculate CEs: $H[Y|V0]$ as shown in the left of Figure 4.17. In sharp contrast, due to having a very little number

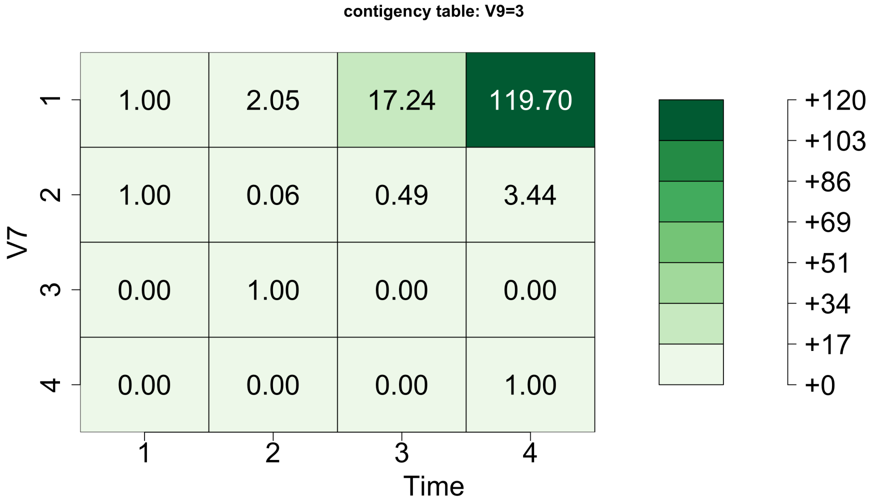


FIGURE 4.16. Contingency table of $C[V7 - vs - T]$ within sub-collection $V9 = 3$.

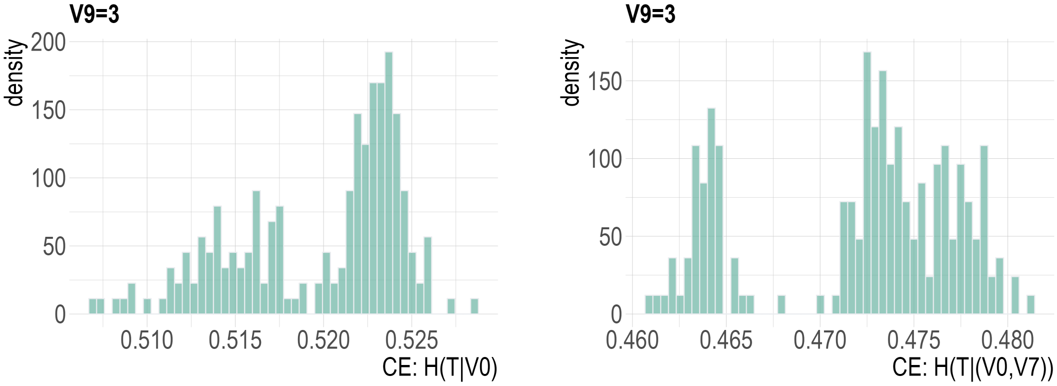
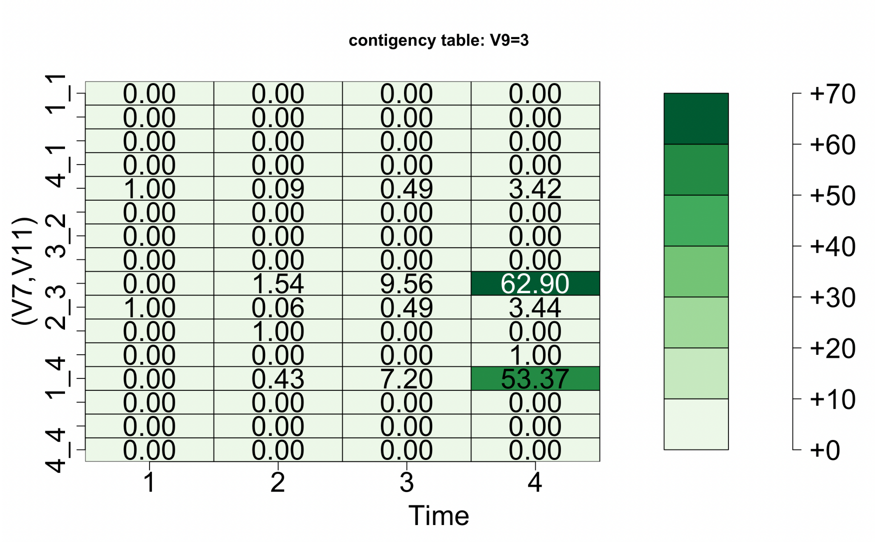


FIGURE 4.17. Null distributions for reliability checks for 1-feature and 2-feature settings: Left: simulated CEs of $H[T|V0]$; Right: simulated CEs of $H[T|(V0, V7)]$ in sub-collection {V9=3} based on 200 simulated $U[0, 1]$ based features.

of uncensored data points in this sub-collection {V9=3}, we see only V7 and V6 have very small p-values. As for the 2-feature setting, we see that no feature-pairs have small p-values according to the right of Figure 4.17 of the null distributions of simulated CEs of $H[T|(V0, V7)]$. However, we still accordingly carry out the task of identifying potential interacting effects for the continuum of our discussion across all sub-collection as follows.



(V7,V11)	1	2	3	4
1_1	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00
4_1	0.00	0.00	0.00	0.00
	1.00	0.09	0.49	3.42
	0.00	0.00	0.00	0.00
3_2	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00
	0.00	1.54	9.56	62.90
2_3	1.00	0.06	0.49	3.44
	0.00	1.00	0.00	0.00
	0.00	0.00	0.00	1.00
1_4	0.00	0.43	7.20	53.37
	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00
4_4	0.00	0.00	0.00	0.00

FIGURE 4.18. Contingency table of $C[(V7, V11) - vs - T]$ within the sub-collection $V9 = 3$.

As for results of 2-feature setting, feature-pairs: {V6, V10}, {V7, V11}, {V10, V11}, achieve ecological effects, but not {V6, V7}. Another interesting observation is that there are no evident interacting effects present in this sub-collection. Therefore, we can conclude a reasonable collection of order-1 major factors: {V7, V11}. The effect of coupling V11 with V7 is seen through the contingency table in Figure 4.18. Their CE-expansion together with CE-expansions with respect to feature-pairs: {V6, V10}, {V7, V11} and {V7, V10}, are reported in Figure 4.19. The degrees of expansions offered by V10 and V11 are relatively small.

{V9=4}: $1(n_o) - vs - 16(n_c)$ In the sub-collection {V9=4}, there is only one data point out of 17 uncensored. From Table 4.14, the CE of T is the smallest among the four sub-collection. Based on CEs of 1-feature and 2-feature settings reported in the table, the rest of the 15 features still offer some extra information beyond the information of {V9=4}. This is another striking aspect of heterogeneity with respect to V9. In sharp contrast, the PH regression model simply doesn't provide any meaningful result.

Based on the results of 1-feature in Table 4.14, the two facts: 1) the top two ranked features are {V12, V14}; 2) V7 is ranked at the bottom, are totally new by being totally different from the previous three sub-collections. From Figure 4.20, the contingency table $C[V12 - vs - T]$ shows the 2nd row with relatively low CEs.

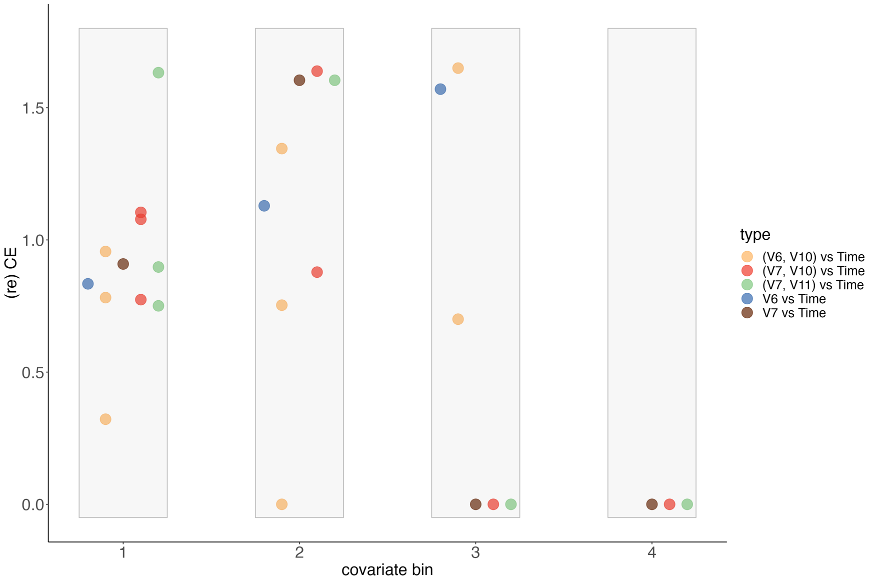


FIGURE 4.19. Heterogeneity of CE-expansion ($V9 = 3$): V7-vs-T (blown dots), V6-vs-T (blue dots), (V6, V10)-vs-T (orange dots), (V7, V10)-vs-T (red dots) and (V7, V11)-vs-T (green dots). The CE of T in sub-collection {V9=3} is equal to 0.5380.

Nonetheless, for the reliability check on CE calculations, we use the same simulation plan and calculate CEs: $H[Y|V0]$ as shown in the right of Figure 4.21. Due to having only one uncensored data point out of 17 in this sub-collection {V9=4}, we see that only V12 has a p-value at the borderline of being significant. As for the 2-feature setting, all feature-pairs have rather big p-values according to the left of Figure 4.21 of the null distributions of simulated CEs of $H[T|(V0, V12)]$. For the continuum of our discussion on interacting effects, we still briefly interpret 2-feature results as follows.

As for the 2-feature results, the feature-pair {V12, V14} achieves the ecological effect, so they can be concurrently present as order-1 major factors. To see the result of such concurrent presence of features V12 and V14, based on Figure 4.22, their contingency table of $C[(V12, V14) - vs - T]$ show 4 rows having relatively low CEs out of 6 non-zero rows. More detailed CEs results are presented in the CE-expansion plots shown in Figure 4.23. Based on these results, we see several univariate- and bivariate categories are qualified for short code-IDs.

Further, we also see feature-pairs: {V11, V12} and {V11, V14}, achieve interacting effects. Thus, if we are to choose a collection of major factors within this sub-collection, the collection of

2-order major factors {{V11, V12}, {V11, V14}} is one reasonable choice. This collection is rather distinct from the three collections chosen in the previous three sub-collection.

1-feature	CE	SCE-drop	p-value(PH)	2-feature	CE	SCE-drop
V12	0.3380	0.0387	1.0000	V12_V14	0.3104	0.0276
V14	0.3582	0.0185	1.0000	V6_V14	0.3131	0.0451
V6	0.3596	0.0171	1.0000	V1_V14	0.3135	0.0446
V1	0.3600	0.0167	1.0000	V12_V13	0.3254	0.0125
V3	0.3660	0.0107	1.0000	V6_V12	0.3261	0.0119
V5	0.3677	0.0090	1.0000	V11_V12	0.3261	0.0119
V10	0.3704	0.0063	1.0000	V11_V14	0.3301	0.0280
V2	0.3715	0.0052	1.0000	V3_V6	0.3323	0.0273
V13	0.3725	0.0042	1.0000	V3_V12	0.3326	0.0053
V4	0.3749	0.0018	1.0000	V1_V12	0.3326	0.0053
V11	0.3753	0.0014	1.0000	V2_V12	0.3333	0.0047
V7	0.3767	0.0000	NA	V3_V13	0.3354	0.0305
V8	0.3767	0.0000	NA	V10_V14	0.3357	0.0225
V15	0.3767	0.0000	NA	V10_V12	0.3361	0.0018
V16	0.3767	0.0000	1.0000	V12_V15	0.3362	0.0017

TABLE 4.14. {V9=4}:Top 15 ranked conditional entropies (CE) and successive CE-drop under 1-feature and 2-feature settings. $n = 17$ with 1 uncensored data point.

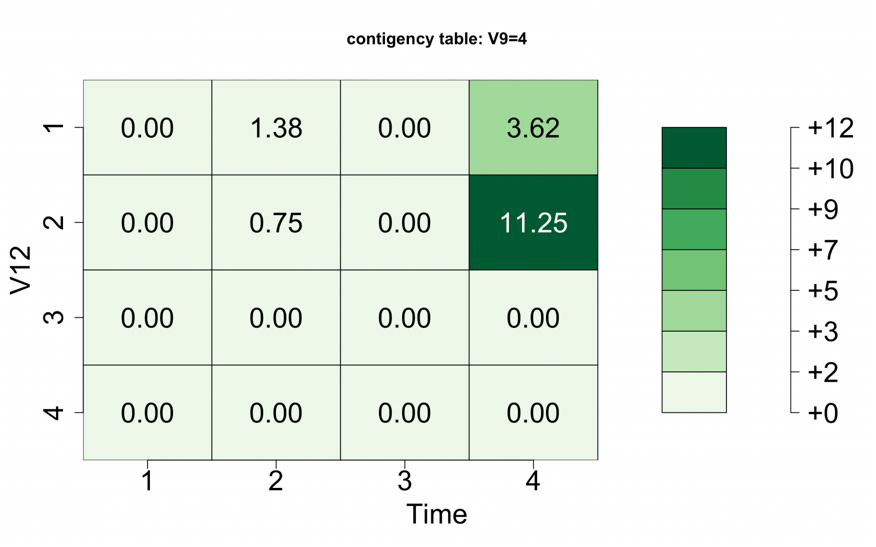


FIGURE 4.20. Contingency table of $C[V12 - vs - T]$ within sub-collection $V9 = 4$.

Overall results of ADNI data analysis from V9 perspective. Among the sub-collections: from {V9=1} to {V9=4}, we see their global distinctions as indicated from the CEDA results

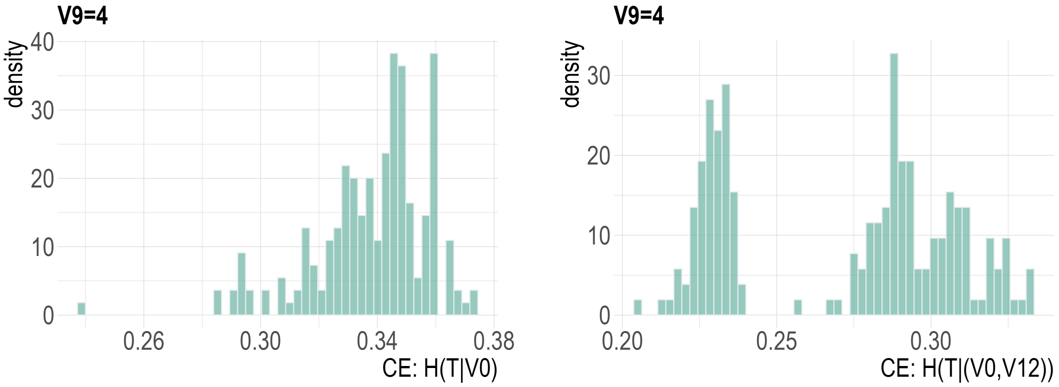
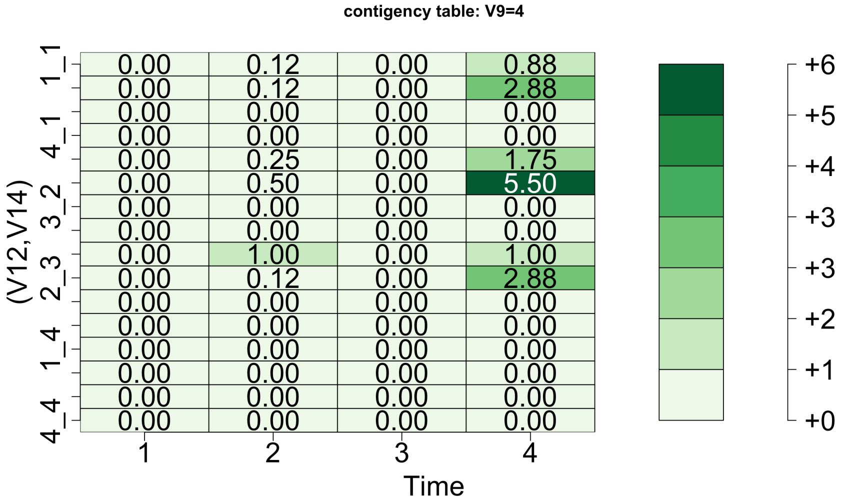


FIGURE 4.21. Null distributions for reliability checks for 1-feature and 2-feature settings: Left: simulated CEs of $H[T|V0]$; Right: simulated CEs of $H[T|(V0,V12)]$ in sub-collection {V9=4} based on 200 simulated $U[0,1]$ based features.



(V12,V14)	1	2	3	4
1_1	0.00	0.12	0.00	0.88
	0.00	0.12	0.00	2.88
	0.00	0.00	0.00	0.00
4_1	0.00	0.00	0.00	0.00
	0.00	0.25	0.00	1.75
	0.00	0.50	0.00	5.50
3_2	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00
	0.00	1.00	0.00	1.00
2_3	0.00	0.12	0.00	2.88
	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00
1_4	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00
4_4	0.00	0.00	0.00	0.00

FIGURE 4.22. Contingency table of $C[(V12,V14)-vs-T]$ within sub-collection $V9=4$.

derived from using the censoring status as a response variable in Section 4.5. The feature V9 is shown to achieve the highest CE-drop as having the largest mutual information with the response variable δ_i. With V9 as one of the natural perspectives for exploring heterogeneity in ADNI data, the first evident heterogeneous pattern is seen from the significant variations of four conditional entropies of T: {0.6545, 1.3183, 0.5380, 0.3767} across the four sub-collections. This evidence

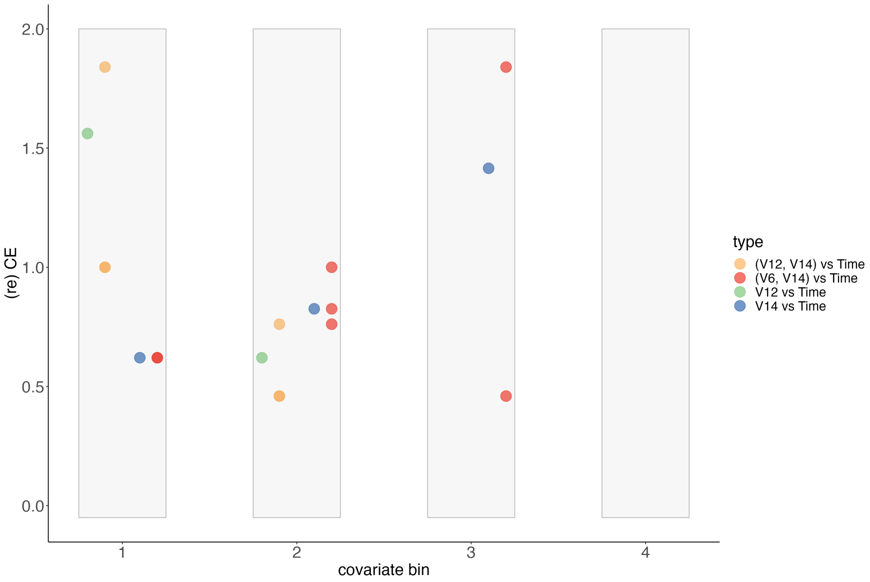


FIGURE 4.23. Heterogeneity of CE-expansions ($V9 = 4$): V12-vs-T (green dots); V14-vs-T (blue dots); (V12, V14)-vs-T (orange dots); (V6, V14)-vs-T (red dots). The CE of T in sub-collection {V9=4} is equal to 0.3767.

reveal structural distinctions among the four compositions of subjects belonging to the four sub-collections. That is, the V9 indeed provides significant amounts of information in sub-collections {V9=1}, {V9=3} and {V9=4}, but not in sub-collection {V9=2}. Such varying results further point to further needed investigations of finding which features or feature-sets can assist V9 to further reduce the uncertainty of T.

The second global heterogeneous pattern is the varying compositions of the top 5 ranked individual features and the four sets of feature-pairs having interacting effects. In particular, the 1st ranked individual features across the four sub-collections are all different. Features V7 and V6 respectively appear as individual 1st ranked features in {V9=1} and {V9=3}. Both features also appear in the top three ranked features across {V9=1} to {V9=3}. That is, V7 or V6 indeed can individually provide some essential amounts of extra information beyond V9, but not both together because this feature-pair doesn't achieve the ecological effect. V4 is ranked 1st in {V9=2}, while V12 is ranked 1st in {V9=4}. In contrast, both features are ranked rather low in {V9=1} and {V9=3}.

It is interesting to note also that the V11 plays a different role across the four sub-collections. It plays the role of candidate of order-1 major factor in {V9=1} and {V9=2}, while uniquely plays the supporting role of having an interacting effect with order-1 major factors in {V9=3} and {V9=4}. To a great extent, V10 also plays more or less the same roles as V11 across these four sub-collections. It is worth mentioning that V3 and V7 reveal their interacting effect in {V9=1}, not in the rest of the three sub-collections.

As for V8, it is a known important feature variable in AD. However, one significant and consistent pattern coming out of our CEDA results is: "V8 is never an obvious order-1 major factor". This observation seemingly means that V8 indeed doesn't provide essential amounts of information about T beyond V9. Is this implication solid? We would see some clues in the next subsection.

Such heterogeneous patterns of global and sub-collection scales are evidently authentic because no man-made assumptions or structures are involved. After recognizing the fact that heterogeneity is inherent in this data set, its implied consequence is also recognized as collected manifestations, each of which rests on all those categories receiving very short code-IDs via one specific perspective. Therefore, the ideal scenario is that all subjects receive a spectrum of short code-IDs derived from many distinct perspectives. This scenario is the topic studied in the Part-II of this paper.

At the end of this section, we conclude that no homogeneous modeling structures could be suitable for this ADNI data set. Its inadequacy demonstrated through the Cox's PH hazard regression model on two scales: overall and the four sub-collections is clear and intuitive. Such intuition is surely not new. It is understood that human aging-related diseases are too complex to be captured by any homogeneity-based modeling structures. Here we further extend our intuition to express that complexity embraced into individual person's disease dynamics can be revealed and presented via the above heterogeneity's manifestation to a great extent as depicted in our ideal scenario.

4.5.6. Heterogeneity w.r.t V8 (ADAS13-bl). We choose to look into heterogeneity through the V9 perspective in the previous section. It is natural to look through V8 as well because V8 ranked 2nd in reducing the uncertainty of δ_i in Section 4.5.3. On the other hand, the associative relation between V8 and V9 is indeed complicated from their contingency table $C[V9 - vs - V8]$ as shown in Figure 4.24. All cells under the diagonal of 4×4 matrix lattice are zeros, while all entry counts on the diagonal are much smaller than entry counts right above the diagonal, which decreases sharply along the other diagonal direction. Such patterns signal multiple non-linear constraints embraced in this contingency table going beyond the simple observation: V9's ordinal categories are arranged in the reverse order of V8's ordinal categories. So it becomes curious to ask whether V8 generates heterogeneity that mirrors heterogeneity pertaining to V9 in the previous subsection.

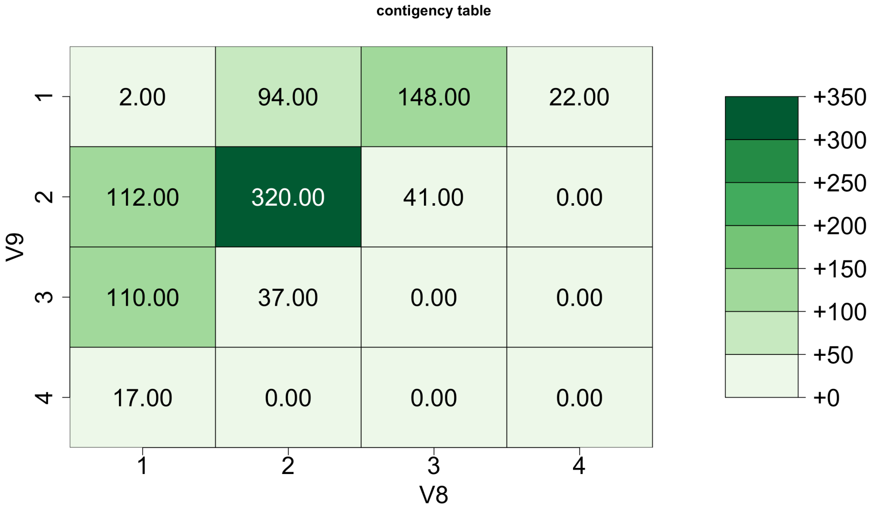


FIGURE 4.24. Contingency table of V9-vs-V8.

The heterogeneity pertaining to V8 would be explored in this subsection with respect to four sub-collections:{V8=1}, {V8=2}, {V8=3} and {V8=4}. It is striking to note that the sub-collection {V8=4} consists of subjects whose survival time from MCI to AD all fall in the category of $T = 1$. This fact is consistent with the fact that higher scores of V8(ADAS13.bl) strongly indicate the likelihood of progressing into a more severe disease [**66**]. Hence, the CE of T is zero, and so are all conditional CEs with respect to all covariate features. This is one striking aspect of heterogeneity that we haven't seen within V9's sub-collections. Therefore, it is worthwhile exploring further the rest of the three sub-collections.

{V8=1}: According to the computed CEs reported in Table 4.15 for the 1-feature setting of sub-collection {V8=1}, V9 is the most dominant covariate feature by achieving a CE-drop that is twice of the CE-drop of the 2nd ranked feature V11. And it is unusual to see V15 ranked 7th, which is lowly ranked across the four sub-collections with respect to V9.

As for the 2-feature setting, feature-pair {V9, V10} is ranked 1st. We know that V9 and V10 are marginally independent. They indeed become conditional dependent given T by achieving the ecological effect. In fact, all 10 feature-pairs of the top 5 ranked features are on the top 15 list and they all achieve ecological effects. Thus, they are natural candidates for order-1 major factors. We also observe the feature-pair: {V3, V6}, achieving the interacting effect.

Hence, if we are to choose a potential collection of major factors, then this collection should be {V7, V9, V10, V11, V4, {V3, V6}}, within this sub-collection {V8=1}.

1-feature	CE	SCE-drop	2-feature	CE	SCE-drop
V9	0.9456	0.0414	V9_V10	0.9126	0.0329
V11	0.9662	0.0207	V6_V9	0.9222	0.0234
V7	0.9669	0.0201	V4_V9	0.9231	0.0224
V4	0.9677	0.0192	V9_V11	0.9244	0.0211
V6	0.9689	0.0181	V11_V15	0.9255	0.0406
V10	0.9706	0.0164	V7_V11	0.9277	0.0384
V15	0.9715	0.0155	V7_V9	0.9282	0.0174
V1	0.9733	0.0137	V1_V11	0.9290	0.0372
V16	0.9772	0.0097	V1_V9	0.9313	0.0142
V13	0.9783	0.0087	V4_V11	0.9314	0.0348
V12	0.9786	0.0083	V9_V12	0.9315	0.0141
V14	0.9811	0.0059	V6_V10	0.9317	0.0372
V3	0.9815	0.0054	V3_V6	0.9318	0.0371
V2	0.9841	0.0029	V9_V15	0.9337	0.0118
V5	0.9848	0.0021	V3_V9	0.9344	0.0112

TABLE 4.15. {V8=1}:Top 15 ranked conditional entropies (CE) and successive CE-drop under 1-feature and 2-feature settings.

{V8=2}: Within the sub-collection {V8=2}, the entropy of T here is much higher than the entropy of T in sub-collection {V8=1}. Based on CEs of 1-feature setting in Table 4.16, the feature V9 is still a dominant factor by having a CE-drop almost twice as large as the CE-drop of the 2nd ranked V7. Even though all involving features are less associated than they are marginal, the number of feature-pairs achieving the ecological effect is much smaller than that in sub-collection {V8=1}. We only find feature-pairs {V9, V10}, {V4, V9} and {V4, V10} achieve the ecological effects. Thus, if we need to conclude with a potential collection of major factors, then this collection is {V9, V4, V10}.

{V8=3}: Based on Table 4.17 for the sub-collection {V8=3}, V11 is ranked the 1st. That is, the 2nd-ranked V9 is no longer a dominant feature in terms of CE-drop. The most striking

1-feature	CE	SCE-drop	2-feature	CE	SCE-drop
V9	1.1884	0.0702	V7_V9	1.1537	0.0346
V7	1.2171	0.0415	V4_V9	1.1592	0.0291
V6	1.2260	0.0327	V6_V9	1.1606	0.0278
V4	1.2364	0.0223	V9_V10	1.0681	0.0202
V11	1.2373	0.0213	V9_V11	1.1715	0.0168
V5	1.2428	0.0158	V9_V15	1.1725	0.0159
V10	1.2451	0.0135	V9_V12	1.1744	0.0139
V12	1.2470	0.0117	V9_V14	1.1746	0.0137
V13	1.2472	0.0114	V9_V16	1.1765	0.0118
V14	1.2506	0.0080	V1_V9	1.1781	0.0103
V16	1.2514	0.0072	V9_V13	1.1804	0.0080
V1	1.2517	0.0069	V3_V9	1.1841.	0.0042
V15	1.2527	0.0059	V2_V9	1.1845	0.0039
V3	1.2553	0.0033	V7_V11	1.1851	0.0320
V2	1.2581	0.0005	V4_V7	1.1903	0.0267

Table 4.16. {V8=2}:Top 15 ranked conditional entropies (CE) and successive CE-drop under 1-feature and 2-feature settings.

observation is that 14 out of the top 15 feature-pairs are having ecological effects, except {V9, V11}. This phenomenal pattern is almost entirely opposite of that of the top 15 feature-pairs observed in the sub-collection {V8=2}. On top of this specific observation, there is only one out of the 15 pairs having interacting effects:{V1, V11}. We conclude that if we are to choose a collection of major factors, this collection is: {{V1, V11}, V7, V6, V4, V14, V3}, which is very distinct to the selected collection in the sub-collections {V8=1} and {V8=2}. This is an evident perspective of heterogeneity contained in this data set.

Overall results of ADNI data analysis from V8 perspective. Across the four sub-collections, the four CEs of T vary significantly: {0.9869, 1.2586, 0.8355, 0.0000}. This is the first evidence of heterogeneity. The feature V9 is the top-ranked in sub-collections: {V8=1} and {V8=2}, but ranked 2nd in {V8=3}. This is the second piece of evidence. The three collections of feature-pairs that achieve the ecological effect or interacting effect in the sense of becoming conditional dependents are very distinct. This is the third piece of evidence. Thus, we conclude that V8 is also a legitimate perspective for heterogeneity in this ADNI data set.

These pieces of evidence of heterogeneity echo the suggestion provided in the review paper [**66**]: the original ADAS-Cog is not an optimal outcome measure for pre-dementia studies. It needed modification. It went on to suggest that the most beneficial modification of ADAS-Cog is tests of

1-feature	CE	SCE-drop	2-feature	CE	SCE-drop
V11	0.7895	0.0461	V7_V11	0.7363	0.0531
V9	0.7926	0.0429	V9_V11	0.7482	0.0412
V7	0.7931	0.0424	V7_V9	0.7483	0.0442
V6	0.8041	0.0315	V6_V11	0.7505	0.0389
V16	0.8099	0.0256	V4_V11	0.7508	0.0386
V4	0.8192	0.0163	V1_V11	0.7519	0.0375
V14	0.8202	0.0153	V7_V16	0.7542	0.0389
V3	0.8226	0.0129	V11_V14	0.7545	0.0349
V15	0.8229	0.0126	V1_V7	0.7553	0.0378
V1	0.8249	0.0106	V11_V12	0.7554	0.0339
V10	0.8251	0.0104	V7_V14	0.7560	0.0371
V2	0.8256	0.0099	V4_V7	0.7562	0.0369
V13	0.8297	0.0058	V3_V11	0.7573	0.0332
V12	0.8324	0.0031	V6_V7	0.7588	0.0342
V5	0.8331	0.0024	V3_V7	0.7599	0.0332

TABLE 4.17. {V8=3}:Top 15 ranked conditional entropies (CE) and successive CE-drop under 1-feature and 2-feature settings.

memory. As it turns out to be V9 in this study. Since, across the first three sub-collections, V9 consistently reveals to offer extra information beyond V8.

Here we reiterate that the particular valuable piece of information offered by V8 is the zero CE in {V8=4}. That is, the 22 subjects' status of being in {V8=4} can precisely point to one and only one category #1 of T. From the same perspective of having very low CEs, the two pieces of information of being {V9=3} and {V9=4}, respectively, are also critical for the 147 and 17 subjects in the two sub-collections. In sharp contrast, this is not the case for the 189 subjects in sub-collection {V8=3}. These pieces of information are parts of the collective heterogeneity embedded within the information content of this ADNI data.

In summary, information pieces are indeed scattered among various perspectives of heterogeneity. This is the reality at least in this ADNI data set. These valuable fragmenting pieces of information need to be collected and systemized. Therefore, no matter whether the data analyzing goal is focused on an understanding of the complex system under study or just for predictive decision-making, such systemizing information pieces is a critical task. For instance, the critical issue facing the 189 subjects in sub-collection {V8=3} is how to more precisely describe these subjects by incorporating which feature variables? That is the chief purpose of suggesting a collection of major factors within sub-collections, with which improvements in describing subjects' characteristics can be potentially derived.

4.6. Conclusions and discussions

In the previous two subsections, we have demonstrated two perspectives of heterogeneity embedded within this data set. There are many other perspectives to be explored to build the data's information content more fully. A version of the data's full information content can be described in the ideal scenario as mentioned in the conclusion of the subsection devoted to heterogeneity from the V9 perspective. In this ideal scenario, each subject is attached to a spectrum of code-IDs from explored different perspectives of heterogeneity. These Code-IDs would tell what are relevant pieces of information pertaining to this subject: some pieces say which features or feature-sets would shed clear lights on which category of this subject's T would fall into with short coding lengths and very low uncertainty, while in contrast, some pieces say which features or feature-sets can only support potential categories of T with high uncertainty. Both kinds of pieces of information and the spectrum of code-IDs would help facilitate a better understanding of AD as a complex disease. At this stage, making such a manifestation of heterogeneity-based full information content is studied and deferred to Part-II of this study.

In this paper as Part-I, when combining results of heterogeneity derived from the V9 and V8 perspectives, we clearly see that V9 provides extra information beyond V8, while V8 doesn't provide extra information beyond V9. However, the results of heterogeneity derived from the four sub-collections via the V8 perspective surely offer to distinguish pieces of information beyond the results of heterogeneity from the four sub-collections via the V9 perspective. Such a discrepancy between information and entropy indeed points to a very important direction for extracting information from data: all relevant perspectives of heterogeneity are worth exploring. Different relevant perspectives of heterogeneity pertaining to different features or feature-sets mount to offer distinct pieces of information with distinct implications. Therefore, the immediate and critical issue for building the ideal scenario is how to exhaustively search for all potentially relevant perspectives of heterogeneity. Explicit resolutions to this issue would need further computational developments, which are also referred to the Part-II. Further, another critical issue arises: after knowing what important pattern information should be harvested from each perspective of heterogeneity, we need to address how to effectively display and systemize valuable, but fragmented information pieces to build a scientific understanding of the AD's prognosis dynamics.

In this paper, we adopt the CEDA paradigm to analyze this ADNI data set of time needed from MCI to an event of AD diagnosis. We first propose a formal testing protocol for checking whether the censoring mechanism is independent of the targeted time-to-event. This contingency-table-based approach is a rather straightforward application of the Re-distribution-to-the-right algorithm. However, to our knowledge, this simple approach has not yet been reported in Survival Analysis literature. After confirming the non-informative censoring mechanism, we compare CEDA results with Cox's PH results on two scales: global and sub-collection. Due to heavy censoring rates, all PH results on both scales deem to be unreliable. Therefore, heterogeneity becomes an urgent and critical issue in Survival analysis. And the immediate impact of the presence of heterogeneity in time-to-event data is that any modeling construct with a global structure, such as the linearity-based structural PH model, has slim chances of producing scientifically valid results.

Further, as we really acknowledge that a complex disease, such as AD, indeed retains very little chance to be fitted well by any global model due to existential heterogeneity, the whole disease progressing process doesn't likely sustain man-made additive effects from different features and feature-sets as typically assumed in statistical modeling. Furthermore, any interacting effects due to any sets of conditional dependent features of various order can not be regressed into any fixed formats preferred just for implementing mathematical or statistical operations. Thus, it is a conservative, robust and realistic way of approaching data analysis by simply admitting that data analysts have acquired no knowledge regarding complex formats of effects of features or feature-sets on global as well as local scales. This way of data analysis is the philosophic basis of the CEDA paradigm.

CHAPTER 5

Analysis of variability of functionals of recombinant protein production trajectories based on limited data

5.1. Introduction

Many biological experiments involve production of certain recombinant molecule over a period of time under different experimental conditions. Thus, the data associated with such experiments are inherently longitudinal. One long-standing problem is to compare these optimum trajectories across different factors or experimental conditions (or treatments), which is a core topic of longitudinal data analysis [**38**, **48**, **65**, **73**]. In most of these studies, the object of interest is typically the expected amount of the ingredient being measured, and one has multiple replicates to accommodate a comprehensive ANOVA (Analysis of Variance) approach to deal with the problem of ascribing effects of various factors.

Indeed, the traditional approach to such inferential questions has been through the application of repeated measures designs [**42**, **93**, **105**]. However, in many real-life lab-based biological experiments, one key constraint is the number of data points or replicates that can be obtained, due to the time, costs and resources associated with completing each condition, particularly in scaling-up experiments.

Furthermore, in many instances, as we discuss below, the key object of interest is not the level of the target molecule itself but some, possibly nonlinear, functionals of the production trajectory. For instance, this functional could be (a) the time it takes for the accumulation of the target molecule to reach a prespecified value (to be referred to as the "optimal time-to-harvest"); (b) the maximum production level (= the maximum of the production trajectory); or (c) the maximum productivity, defined as the maximum of the amount divided by time over the duration of the experiment.

5.1.1. Scientific Context. As a good example, butyrylcholinesterase (BChE) circulating in human blood plasma is a tetrameric hydrolase enzyme that can be potentially used as a prophylactic and/or therapeutic factor against organophosphorus nerve agent poisoning [**74**]. However, the use of purified BChE from human blood plasma in clinical stages is limited by its cost, which is estimated to be $20,000 per 400 mg dose [**3**]. Thus, recombinant human BChE (rBChE) has been developed in several host expression systems, including transgenic rice cell suspension cultures, to be used as an alternative source of BChE.

Our lab developed metabolically-regulated transgenic rice cell suspensions under the rice alpha amylase 3D (RAmy3D) promoter to produce rice-made recombinant human BChE (rrBChE) [**19**, **76**]. In nature, the RAmy3D promoter in rice cells derived from rice seed, is suppressed in a sugar-rich environment but activated in a sugar-starved environment [**49**, **50**, **99**]. In other words, the RAmy3D promoter-based transgenic rice cell suspensions are grown in a sugar-rich medium for production and transferred into sugar-free medium for rrBChE production.

Biological experiments of the kind described above are both time-consuming and expensive. A major challenge of growing plant cell suspension cultures is the slow growth rate of plant cells compared to microbial and mammalian cells. For example, it takes 6–7 days for transgenic rice cells to reach mid-to-late exponential growth phase, followed by the medium exchange to replace spent growth medium with sugar-free medium, and another 4–5 days post-induction for rrBChE expression [**19**, **76**]. In other words, the cultivation time of transgenic rice cell suspensions in a batch culture is 10–12 days.

When it comes to an experiment with several factors or conditions, the number of bioreactor replicates is likely to be restricted due to time of cultivation and limited equipment. Therefore, there might be difficulties when interpreting the data or choosing the optimal sets of conditions. Given the cost, the information delivered by the data is crucial, and thus we would like to understand the data in a more comprehensive way by developing statistical methods. For instance, it will be interesting and meaningful to characterize the production curves over time, particularly when the measurement time points are limited in practice.

We are able to predict production quantities at any experimental time points, in addition to those at observed time points. Furthermore, statistical inference will be useful to measure and indicate the factors' impact on the difference among multiple experimental conditions in terms of certain metrics. Thus, in the aforementioned rrBChE study, we would like to provide a robust and effective statistical approach as a validated way to interpret the data better and address experimental questions through a statistical framework. For example, we can use inference procedures to determine and compare varying metrics, such as the “optimal time-to-harvest” of each factor based on certain levels of statistical significance, which will indicate the effect of factors behind the limited data.

Therefore, in this study, we employ novel statistical approaches to tackle limited data to build trajectory models using previously reported bioreactor data [**76**] to predict outcomes of interest, such as the “optimal time-to-harvest”, maximum rrBChE production level and maximum productivity. In addition, estimating the trajectory of the production level is a part of Quality by Design (QbD) [**91**] that is essential in biomanufacturing where a computationally feasible statistical method is involved in modeling based on available data.

5.1.2. Statistical Challenges and Contributions. Analysis of the variability of functionals of protein production trajectories across different experimental conditions presents several novel statistical challenges. One key requirement is to ensure that the underlying production trajectories are monotonic (i.e., either increasing or decreasing functions of time), without which some of the quantities of interest are not even properly defined. At the same time, due to the limited number of data points at which these trajectories are typically measured, it is imperative to borrow information across trajectories in order to ensure that we have sufficient degrees of freedom left for comparing the parameters across the factors.

Another challenge is that, due to both the limited number of data points and the restrictions imposed by the monotonicity of production trajectories, any statistical inference procedure that directly relies on large sample theory will have limited accuracy or may be misleading. Moreover, since some of the parameters (functionals of the production trajectories) or process metrics of interest are nonlinear, the standard ANOVA framework that relies on the linear model theory does not apply.

In this paper, we primarily focus on comparing the equality of the parameters by means of simultaneous pairwise comparisons, which can be formulated as a multiple hypothesis testing problem. Thus, the key statistical challenge is to develop a methodology that (a) ensures monotonicity of the fitted trajectories and (b) can handle simultaneous inference for arbitrary functionals of the trajectories with a limited amount of data.

In order to address these challenges, we adopt the following three-pronged approach. First, following the ideas in [**58**], we model the production trajectories by representing them in a B-spline basis and incorporate the monotonicity constraint by imposing linear inequality constraints on the B-spline coefficients.

We fit the trajectories by using a constrained least squares regression procedure that is implemented through a quadratic programming approach. Next, for statistical inference on the parameters of interest, we use *bootstrap*, or resampling procedures. We compare the efficacies of several different versions of bootstrap, namely, the residual bootstrap, parametric bootstrap and nonparametric bootstrap.

Finally, since we conduct simultaneous inference involving many pairwise comparisons, we adopt a method for imparting control on the *false discovery rate* (i.e., the fraction of false detections) while constructing the simultaneous confidence intervals involving many parameters, using a technique developed in [**10**]. In summary, we provide a comprehensive framework for simultaneous statistical inference on several process metrics that are functionals of biochemical production (or growth) trajectories, based on fairly limited amounts of data, with empirical validity.

5.1.3. Goals of the Study. For the biological experimental study, the goal is to develop an effective and efficient system that is able to scale-up the production of rrBChE given the costs and limited resources. From the statistical side, one of the goals of this study is to analyze the variability of production trajectories for a limited data set of the recombinant protein production by using rrBChE as a model study.

Another goal is to use statistical approaches to determine the optimal time to harvest a recombinant protein during a protein production process. A further goal is to compare across different bootstrap procedures for their relative effectiveness in terms of statistical inference when the data are limited. The last goal is achieved through an extensive numerical simulation study mimicking the recombinant protein production experiments.

5.2. Methods and Materials

5.2.1. Data Collection Method. A 5 L bioreactor (BioFlo 3000, formerly New Brunswick Scientific, Eppendorf Inc., Hauppauge, NY, USA) was used to study the production of rrBChE under eight different conditions as previously described [**76**] and summarized in Table 5.1. In brief, the effects of dissolved oxygen (DO) were conducted in factors (runs) A–E using a two-stage batch culture (the medium exchange was performed to replace spent sugar-rich medium with sugar-free medium to induce the promoter).

Experiment #	%DO during growth phase	%DO during induction phase	Media exchange
A	40	10	Yes
B	40	20	Yes
C	40	30	Yes
D	40	40	Yes
E	40	Uncontrolled	Yes
F	40	40	No
G	Uncontrolled	Uncontrolled	No
H[1]	Uncontrolled	Uncontrolled	No

[1] Initial sucrose concentration in the medium in run H was reduced to 15 g/L instead of 30 g/L used in other runs. DO, dissolved oxygen.

Table 5.1. Conditions used in bioreactor runs A–H when agitation rate and temperature were maintained at 75 rpm and 27^o C, respectively, in all runs. The aeration rate was maintained at 0.2 vvm (volume of sparged gas per working volume per minute) in runs A–F but 0.2–0.4 vvm in runs G and H (reproduced from [**76**])

Factors F, G and H were operated in single-stage batch culture (no medium exchange; production was simply induced by sugar depletion from cellular uptake) with or without controlling DO and using 50% of the usual initial sucrose concentration during the growth phase. For each factor, samples were taken every day during days 0–5 post induction (dpi) to quantify the rrBChE activity in the cell extract and culture medium using a modified Ellman assay [**30**] and assuming a specific activity of 260 U/mg crude rrBChE to convert the activity to the rrBChE production level (μ g/g fresh weight of rice cells) [**76**].

5.2.2. Modeling Production Trajectories. There were $I \geq 2$ factors (or treatments), and each treatment was applied to several independently chosen experimental units (bioreactors). Further, the response (e.g., the rrBChE concentration in the bioreactor) was measured at observation times $0 < t_{i1} < \cdots < t_{iJ} = T$, say, for $J \geq 2$ (this allows the observation times to be different for different factors). Let us denote the mean response curve (at time $t \geq 0$) corresponding to the i-th factor as $\mu_i(\cdot)$. We assumed that $\mu_i(\cdot)$ is a monotonically increasing function of time, over the observation time window $[0, T]$.

For simplicity as well as statistical efficiency, we assumed a *balanced experimental design*—that is, the sample size at each observational time was the same for every treatment. The measurement process is destructive, and thus, for any particular experimental unit, we only have one measurement, at the time of sampling the bioreactor. Therefore, to obtain reasonably accurate measurement for the whole trajectory, we required replicates (i.e., multiple experimental units) for each time t_{ij} and each treatment i.

Let n denote the number of replicates assigned to each combination (i, j), which corresponds to a balanced design. Note that we allow $n = 1$ since, in practice, only limited data are available particularly in certain biological experiments. We denote the response from the k-th experimental unit, in the i-th factor group, measured at time t_{ij} , to be Y_{ijk}.

$$Y_{ijk} = \mu_i(t_{ij}) + \epsilon_{ijk}, \quad j = 1, \ldots, J; k = 1, \ldots, n; i = 1, \ldots, I. \tag{5.1}$$

where ϵ_{ijk} are independent random variables with mean 0 and unknown, common variance $\sigma^2 > 0$. In practice, we may allow the number of time points J to depend on index i as well.

We used a basis representation approach for modeling the mean trajectories μ_i. In particular, we used cubic B-spline basis functions [**23**] for representing the functions. For each i, assuming L cubic B-spline basis functions are used to $\mu_i(t)$, we can write

$$\mu_i(t) = \Sigma_{l=1}^{L} \alpha_{il} B_l(t), \quad i = 1, \ldots, I \tag{5.2}$$

where $\{\alpha_{il}\}$ are the basis coefficients. The number L of basis functions used to model the growth/production trajectories is a user-specified positive integer that controls the degree of complexity of the trajectories, with larger values allowing for more complex shapes. In practice, L may be determined by utilizing data from pilot studies through a cross-validated linear regression procedure, which involves setting aside a random subset of the data (referred to as "validation data") and using it to compare the prediction errors of the fitted trajectories corresponding to different values of L based on the "training data".

A great advantage of the spline representation is that, since the function $B_l(\cdot)$ is non-negative, the curve μ_i is nonnegative provided the coefficient α_{il} is so . A more significant advantage, from the point of view of modeling "production curves" of the type considered here, is that the condition that $\mu_i(t)$ is non-decreasing in t can be imposed by simply requiring that

$$\alpha_{i(l+1)} \geq \alpha_{il} \quad \text{for} \quad l = 1, \ldots, L-1 \tag{5.3}$$

for all i [**58, 68**]. This is a reasonable assumption for batch production of a recombinant protein if the protein is stable in the culture medium (e.g., there is no consumption and/or degradation of the product but simply accumulation due to production). Note that this model is a simplified version of a more general ANOVA framework that enables quantifying possible interactions between the treatments and time. In Section 5.4, we discuss this possible extension to a two-factor ANOVA with potential interactions, and the associated linear constraints.

If we ignore the inequality constraints (5.3), the linear model given by (5.1), (5.2) can be fitted through an ordinary least squares procedure. The resulting estimate of $\mu_i(t)$ for any t will be *unbiased* (i.e., the average of the estimates across all possible samples equals the true value of the parameter) and will have an approximate Gaussian distribution for reasonably large values of n. In this case, we can rely on the large sample theory for statistical inference on the parameters of interest.

However, in the applications considered here, we need to consider the monotonicity constraints (5.3) for modeling the production/growth trajectories. The least squares approach to fitting the model given by (5.1), (5.2), and (5.3) results in a **quadratic programming problem**. Though such estimates guarantee the monotonicity of the mean response curve, the estimates of $\mu_i(t)$ incurs small but non-negligible biases, particularly when the sample sizes are small. The monotonicity constraints on the mean trajectories and limited number of replicates both limit the application of classical large sample theory in dealing with the inference problem. Therefore, we develop a resampling-based strategy for statistical inference.

5.2.3. Key Parameters of Interest. We present the mathematical formulation of the inferential questions associated with the parameters of interest mentioned earlier.

(1) **Optimal time-to-harvest:** $\theta_i = \min\{t : \mu_i(t) = c\}$ for $i = 1, \ldots, I$, where c is the prespecified cut-off level. The corresponding null hypothesis representing no factor effect on the "optimal time-to-harvest" is:

$$\theta_1 = \cdots = \theta_I. \tag{5.4}$$

With $\theta_i = \mu_i^{-1}(c)$ for some given cut-off level c, we are interested in testing the one-sided null hypotheses of the form $\theta_1 \geq s_1, \ldots, \theta_I \geq s_I$ (here, the times $s_1, \ldots, s_I$ need not be equal). These hypotheses translate to the linear inequality constraints:

$$\mu_i(s_i) \leq c \text{ for all } i = 1, \ldots, I \tag{5.5}$$

Notice that, $\theta_1 = \cdots = \theta_I$ is not a linear constraint.

- However, the null hypothesis $\theta_1 = \cdots = \theta_I = s_0$ can be translated to the equality constraints:

$$\mu_1(s_0) = \cdots = \mu_I(s_0) = c. \tag{5.6}$$

- A composite null hypothesis of the form $s_L \leq \theta_i \leq s_U$ for all i can also be translated into linear inequality constraints

$$\mu_i(s_L) \leq c \leq \mu_i(s_U) \tag{5.7}$$

for all i.

(2) **Maximum production:** $\tau_i = \max_t \mu_i(t)$ for $i = 1, \ldots, I$. The regarding null hypotheses are

$$\tau_1 = \cdots = \tau_I. \tag{5.8}$$

or equivalently

$$\mu_1(T_{\max}) = \mu_2(T_{\max}) = \cdots = \mu_I(T_{\max}). \tag{5.9}$$

where $T_{\max}$ is the largest time point during the experiment.

(3) **Maximum "unweighted" productivity:** $\psi_i := \max_t h_i^u(t)$, for $i = 1, \ldots, I$, where $h_i^u(t) = \frac{\mu_i(t)}{t+T_i^{(c)}}$ represents "unweighted" productivity of the i-th factor and $T_i^{(c)}$ is the number of days of cultivation before the induction. Here, "unweighted" means that we do not consider the rice cell fresh or dry weight. The corresponding null hypotheses are

$$\psi_1 = \cdots = \psi_I \tag{5.10}$$

or equivalently

$$\max_t h_1(t) = \cdots = \max_t h_I(t). \tag{5.11}$$

(4) **Optimum stopping time:** Suppose the decision to harvest is taken based on the relative gradient of $\mu_i(t)$ (or, gradient of $\log \mu_i(t)$). Let $\gamma_i = \min\{t \geq T_b : \mu_i'(t)/\mu_i(t) \leq \tilde{c}\}$ where T_b is some constant baseline time and $\tilde{c}$ is a gradient threshold. We may be interested in testing hypotheses of the form $\gamma_i > s_0$ for $i = 1, \ldots, I$ where time s_0 is treated as the same for all i for simplicity. Since $\mu_i'(t)/\mu_i(t)$ is not necessarily monotonic, this cannot be easily reduced to a simple set of inequality constraints. However, we may discretize time to a grid of the form $T_b = T_1 < \cdots < T_m = s_0$ and then consider the slightly relaxed form of the hypothesis

$$\mu_i'(T_j) > \tilde{c}\mu_i(T_j) \quad \text{for } j = 1, \ldots, m, \text{ for all } i. \tag{5.12}$$

In practice, m needs to be small for the feasibility of the optimization problem.

In a more general sense, if our interest is in testing for equality of Θ_i where Θ_i is a linear functional of $\mu_i(t)$, then we can design a procedure that imposes a less restrictive null hypothesis. This applies, for example, to the case when the maximum production level is the quantity of interest, since, under the monotonicity of μ_i, this is simply the value at the maximum time point.

Specifically, suppose $\Theta_i = A(\mu_i)$ where A is a linear operator taking a scalar value, then Θ_i can be expressed as $\sum_l r_l \alpha_{il}$ where r_l are some known coefficients and α_{il} represents the l-th spline coefficient of $\mu_i(t) = \sum_l \alpha_{il} B_l(t)$.

- Example 1: if $\Theta_i = \mu_i(T)$, then $r_l = B_l(T)$.
- Example 2: if $\Theta_i = \int \mu_i(t)dt$, then $r_l = \int B_l(t)dt$.

5.2.4. Statistical Inference Using Bootstrap. We mentioned earlier that the estimate $\hat{\mu}_i(t)$ obtained by imposing constraints (5.3) or (5.16) is not unbiased. What matters more, however, is that we have such a limited number of replicates that we cannot rely on *large sample theory* for making inferences. In view of this, it is imperative to adopt an inferential framework that does not depend too heavily either on the model assumptions, the methodology, or indeed the sample sizes.

Therefore, as a possible alternative, we propose to conduct hypothesis tests or construct confidence intervals for treatment-specific parameters. to be generically denoted by Θ_i (for i-th treatment), by making use of an appropriate resampling procedure. Below, we first describe the different types of resampling strategies that can be employed, depending on the data available. This is followed by specific choices for constructing confidence intervals or performing hypothesis tests involving one or many parameters.

Resampling Strategies

Depending on the structure and amount of available data, we have several resampling or bootstrap strategies for computing the confidence intervals and performing hypothesis tests for the parameters of interest. The detailed procedures can be found in Section S.1 in the Supplementary Materials.

Nonparametric bootstrap with replicates: This variant of resampling can be used when the number of replicates n is relatively but not extremely small, and thus we are able to construct the confidence interval for one parameter (e.g., the difference between a fixed pair of treatments). In some instances, specifically when the null hypotheses is formed by imposing linear equality constraints on the parameters, *bootstrap sampling distribution of the test statistic* (i.e., the histogram of the statistic used for testing the hypothesis, computed from the resampled data), under the null hypothesis can be calculated through a modified version of the nonparametric bootstrap procedure.

This is done by replacing the original data with a set of "surrogate bootstrap data" that incorporates the constraints imposed by the null hypothesis. This enables an efficient computation of the p-values for the hypothesis being tested.

Residual bootstrap with or without replicates: This can be implemented whether there are replicates or not. The core idea here is to resample the residuals from fitting the model to the data. Note, however, that residual bootstrap is not effective particularly when the number of factors is small (say, if $I = 3$, then the residual bootstrap method is not a good option since resampling does not capture the variability adequately as there are too few measurements to represent the true scale of variability of the data). Therefore, in practice, one needs to make a choice of bootstrap procedures based on the design of the experiment.

Parametric bootstrap: As an alternative to nonparametric or residual-based bootstrap, one can use the parametric bootstrap method if the model assumptions either can be validated (say, based on preliminary data) or if no significant departure from these is expected. As the variance of the noise is unknown, sample variance is typically used as a surrogate. When the number of replicates is small (e.g., $n = 1$), and there are only a small number of treatments, neither nonparametric bootstrap nor residual bootstrap are feasible resampling strategies. In this challenging setting, a parametric bootstrap approach can still be used only if there is prior information on variability (see Section 5.3.2 for more details).

In the proposed parametric bootstrap procedure, the observational noise is assumed to follow either a Gaussian distribution or a t distribution. For application to the rrBChE data, we use a scaled t-distribution with relatively low degrees of freedom, which allows for extreme values, thereby, reflecting the variability in the real data more effectively (Figure 5.1). By comparing the confidence intervals, we can see that the results in this real data application are similar regardless of the type of noise, thus, affirming a degree of robustness of the proposed method.

Inference for a Single Parameter

We discuss more details about constructing confidence intervals and p-value computation.

Construction of confidence intervals: We propose two methods to construct the bootstrap confidence intervals:

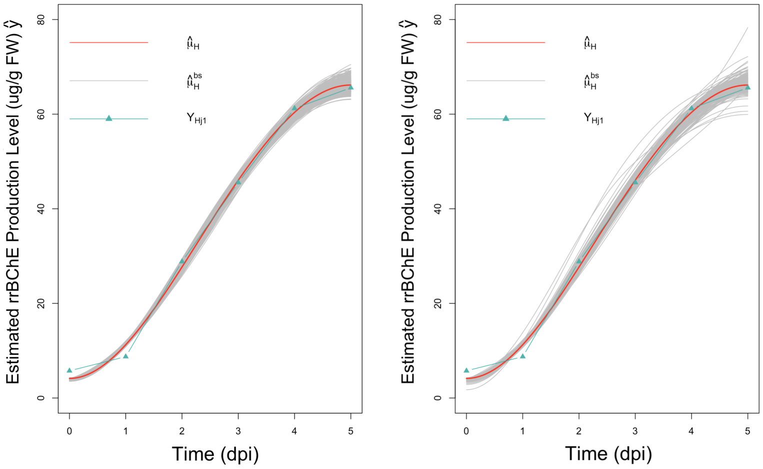


FIGURE 5.1. Observations, fitted curves and 500 bootstrap fitted curves for factor H; **left**: assuming normal noise; and **right**: assuming t-distributed noise.

(1) *Percentile bootstrap confidence intervals*: We obtain percentile bootstrap confidence intervals for Θ_i and $\Theta_i - \Theta_j$ based on B bootstrap estimates of these parameters. The intervals are constructed by using appropriate quantiles of the bootstrap estimates $\{\hat{\Theta}_i^{*b}\}_{b=1}^{B}$ and $\{\hat{\Theta}_i^{*b} - \hat{\Theta}_j^{*b}\}_{b=1}^{B}$, respectively.

(2) *Bias-corrected and accelerated bootstrap interval BC_a:* The percentile bootstrap confidence interval is only first-order accurate. Additionally, it does not correct for skewness of the sampling distribution. To address this, we use bias-corrected and accelerated bootstrap intervals [**27**], denoted by BC_a, which is not only second-order accurate but also corrects for the skewness in the sampling distribution.

Computation of p-value associated with tests of hypothesis: We propose two methods of obtaining approximate p-values using bootstrap. One is the percentile-t bootstrap, a general procedure, while the other one is a special case where the probability distribution of the test statistic under the null hypothesis can be calculated. For a general description of different types of bootstrap procedures and their theoretical validity, one may refer to [**22**].

p-value computation by percentile-t bootstrap or percentile bootstrap: The key idea behind computation of the p-values is to use the correspondence between hypothesis testing and confidence intervals. Specifically, the p-value is equal to η^* where η^* is the largest value of η such that the

$100(1-\eta)\%$ confidence interval contains the value of the parameter specified under the null hypothesis. We can make use of the percentile-t bootstrap procedure for constructing the confidence intervals for the parameters of interest for any given confidence level η and then "invert" these, as described above, to approximate the p-value. Alternatively, we may use the percentile bootstrap procedure (arguably less accurate) instead of percentile-t bootstrap, to compute the p-values, which incurs lower computational costs.

p-value computation under the null distribution: Instead of the indirect approach that relies on construction of bootstrap confidence intervals, in some instances, we can use a modified form of nonparametric bootstrap that enables generating samples under the distribution specified by the null hypothesis. Step 5 in Section S.2 indicates the method of computing p-values under this setup.

The key to successful application of this strategy is the ability to generate surrogate bootstrap data from the null distribution. This is feasible for example when the null hypotheses are specified by linear equality constraints on the parameters. It is reassuring that, for the real data analysis, where, due to the structure of the problem, we have two different ways of computing the p-values, viz., by inverting the bootstrap confidence intervals or by surrogate bootstrap data generated under the null hypothesis, the two versions of p-values are similar.

Simultaneous Inference and Adjusting p-Values

When we perform simultaneous tests for multiple hypotheses of the form $H_0 : \Theta_i = \Theta_{i'}$ vs. $H_1 : \Theta_i \neq \Theta_{i'}$ for several pairs of treatments $1 \leq i < i' \leq I$, in order to control the *familywise type I error rate* (i.e., the probability of rejecting any of the null hypotheses incorrectly), it becomes necessary to adjust the level of significance of each individual test. This can be achieved by making use of the *Bonferroni procedure* or a *False Discovery Rate (FDR) control* procedure [**28**]. In the present setting, this requires computation of the p-values for each individual test using any one of the procedures described above, as appropriate.

Once we obtain the bootstrap p-values for the different tests, the *Benjamini–Hochberg (BH) procedure* for FDR control is used to first determine the significance level for each pairwise test for a given level of familywise significance. With this, we can adjust the confidence levels of the confidence intervals (*False Coverage-Statement Rate* (FCR)-Adjusted BH-Selected CIs) for the parameters accordingly [**10**]. We describe these procedures in the Section S.3 of the Supplementary Material.

5.3. Results

5.3.1. Simulation Study. In this subsection, we present a simulation study illustrating the effectiveness of the proposed bootstrap-based inference procedures and accuracy of the corresponding confidence intervals. This numerical simulation also allows us to make a comparison among the different bootstrap procedures.

Settings: We assume the number of factors $I = 3$, and the time interval is $(0, 9]$ with number of time points $J = 9$. All factors share the same time points $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. We assume there are $L = 5$ basis functions for cubic B-splines with equally spaced knots. For each time point, we have $n = 5$ replicates. Suppose $Y_{ijk} = \mu_i(t_{ij}) + \epsilon_{ijk}$, where the noise level is $\sigma_j^2 \in (1, 3.5)$. Our main interest is the "optimal time-to-harvest" θ_i.

The simulated data and estimation by the standard least square procedure (without constraints) and quadratic programming framework (with constraints) are shown in Figure 5.2. Though the estimations using two methods are similar in our case, there is a difference. The estimated curve by the least square procedure for factor 3 shows a decreasing pattern at the last time point, while the one with constraints remains non-decreasing (Figure A.1) Since we are interested in estimating the growth curve and "optimal time-to-harvest", how we fit the data matters. In addition, using the standard least square procedure may result in oscillations in estimation. The "optimal time-to-harvest" parameter θ_i is of interest with a pre-specified level $c = 10.3$ (Figure 5.3), and we use the residual (nonparametric) bootstrap method for inference. We compute both percentile bootstrap confidence intervals and bias-corrected and accelerated bootstrap interval (BC_a) since the bootstrap sampling distributions involving θ_2 are skewed (Table 5.2). Both types of intervals suggest that θ_1 and θ_2 are significantly different from θ_3; however, we can see that BC_a is slightly better than the general ones in terms of the length of intervals.

5.3.2. Analysis of rrBChE Data. The rrBChE data is available at Dryad [**71**]. We have $I = 8$ factors labeled A to H and primarily focus on the protein production level (ug rrBChE/g FW rice cells) after sugar induction. We have $J = 6$ time points (days post-induction, dpi). Although one potential issue is that we only have one replicate ($n = 1$) at each time for each factor,

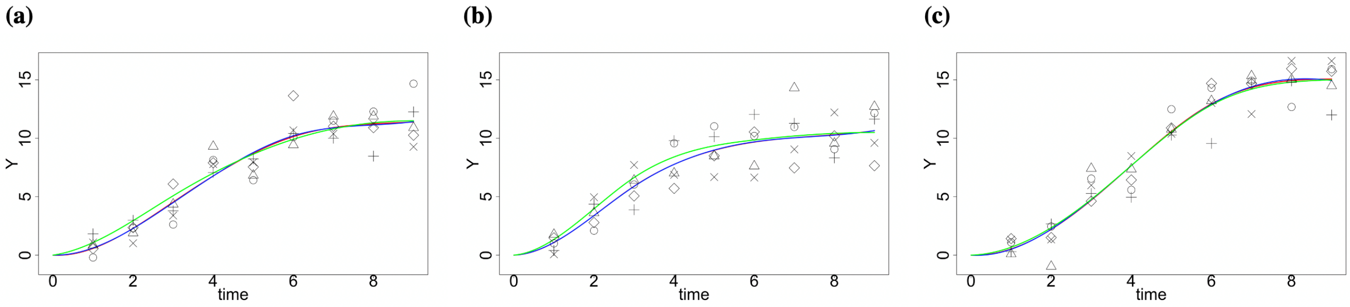


FIGURE 5.2. Simulated data and estimation by the standard least square procedure and quadratic programming framework for different factors; true data are denoted by different types of points; the true curve $\mu(t)$ is marked in green; fitted curves with linear constraints are in red, while the fitted curves by standard least square procedure are in blue (**a**) Factor 1, (**b**) Factor 2 and (**c**) Factor 3.

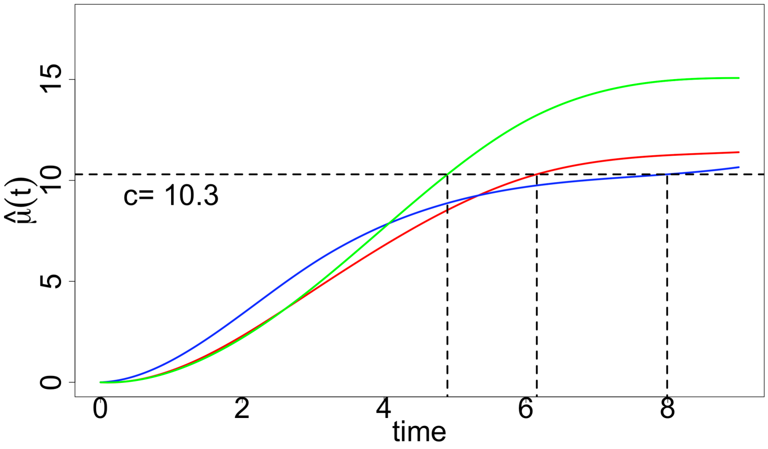


FIGURE 5.3. Fitted curves estimated with constraints (factors 1–3 are in red, blue and green, respectively); $\hat{\theta}_i$ given by a specified level ($c = 10.3$) is indicated by vertical lines .

our framework is able to handle this and make appropriate inferences. The observed data and estimations are shown in Figure 5.4.

“Optimal Time-to-Harvest” θ_i and “Optimum Stopping Time” γ_i

With the pre-specified level $c = 40$ (μg/g FW), we obtain the estimates $\hat{\theta}_i$ and make inferences by using the parametric bootstrap method (Figure 5.5). For the “optimum stopping time” γ_i, we set the level $\tilde{c} = 0.1$ and repeat the procedures. We compute the corresponding confidence intervals by using normal noise and t-distributed noise separately. We found that the results were

		CI Lower	CI Upper	CI Length	True	Mean	sd
Percentile Bootstrap CI	$\hat{\theta}_1^{bs}$	5.820	7.027	1.207	6.378	6.320	0.327
	$\hat{\theta}_2^{bs}$	5.892	9.000	3.108	7.387	7.739	0.995
	$\hat{\theta}_3^{bs}$	4.730	5.162	0.432	4.910	4.917	0.113
	$\hat{\theta}_1^{bs} - \hat{\theta}_2^{bs}$	-3.063	0.568	3.631	$\mathbf{-1.009}$	-1.418	1.046
	$\hat{\theta}_2^{bs} - \hat{\theta}_3^{bs}$	0.937	4.198	3.261	**2.477**	2.822	1.002
	$\hat{\theta}_3^{bs} - \hat{\theta}_1^{bs}$	-2.171	-0.820	1.351	$\mathbf{-1.468}$	-1.403	0.350
BC_a	$\hat{\theta}_1^{bs}$	5.640	6.568	0.928			
	$\hat{\theta}_2^{bs}$	6.036	9.000	2.964			
	$\hat{\theta}_3^{bs}$	4.695	5.126	0.431			
	$\hat{\theta}_1^{bs} - \hat{\theta}_2^{bs}$	-3.189	0.108	3.297			
	$\hat{\theta}_2^{bs} - \hat{\theta}_3^{bs}$	1.117	4.252	3.135			
	$\hat{\theta}_3^{bs} - \hat{\theta}_1^{bs}$	-1.766	-0.525	1.241			

TABLE 5.2. The Percentile Bootstrap Confidence Interval (CI) and bias-corrected and accelerated bootstrap interval (BC_a) for $c = 10.3$. The significance against $H_0 : \theta_i = \theta_j$ is in red. The true pairwise differences are in bold.

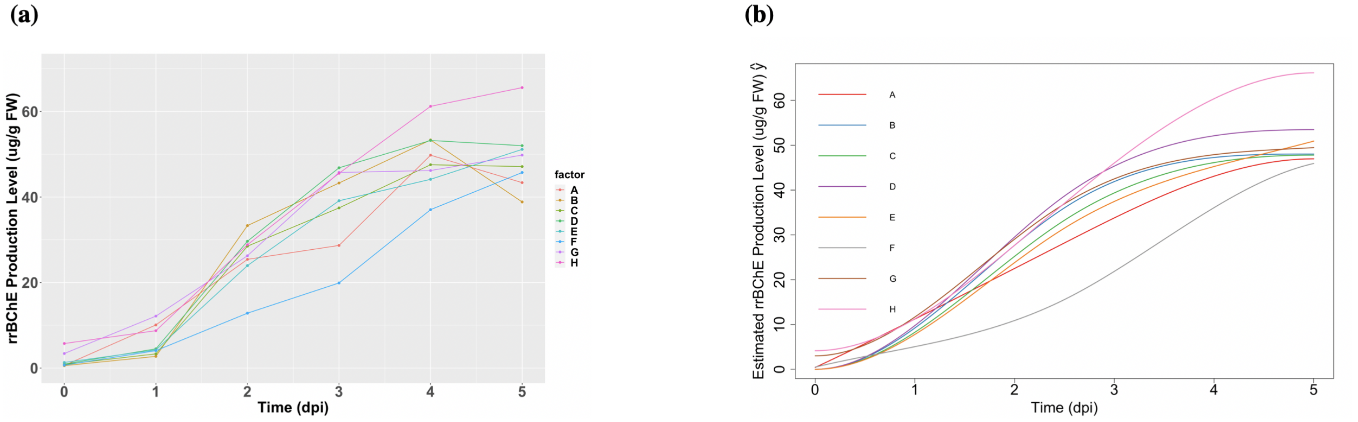


FIGURE 5.4. (**a**): The observed cell-associated rrBChE production levels in factors A to H. (**b**): Fitted curves with monotonicity in factors A to H constraints.

not sensitive to the type of noise (normal or t-distributed) that we are used. More details about the related confidence intervals can be found in Tables A.5 and A.6 in the Supplementary Materials.

Simultaneous Inference—Maximum Production Level τ_i

The hypotheses are:

$$H_0 : \tau_i = \tau_{i'} \qquad \text{vs.} \qquad H_a : \tau_i \neq \tau_{i'}$$

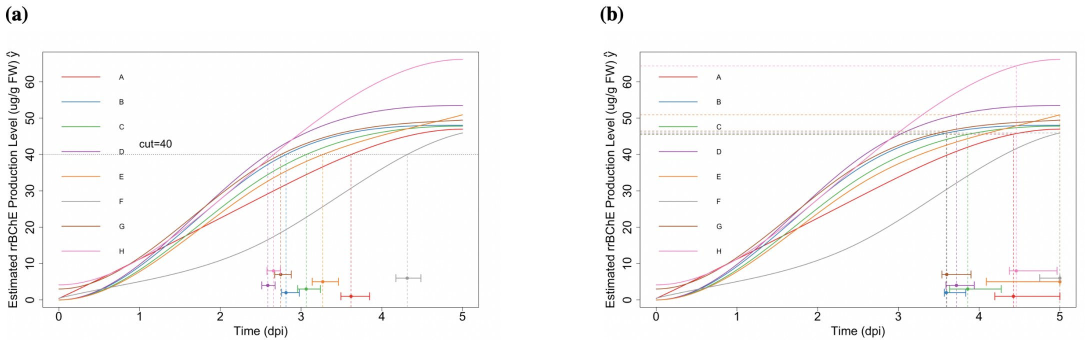


FIGURE 5.5. (**a**): Fitted curves with specified level (cut-off $c = 40$ (μg/g FW)); bootstrap confidence intervals of θ_i are also shown (using t-distributed noise). (**b**): Fitted curves with specified level ($\tilde{c} = 0.1$); bootstrap confidence intervals of γ_i are also shown (using t-distributed noise).

where $\tau_i = \max_t \mu_i(t)$ for $i = 1, \ldots, I$. We use the Benjamini–Hochberg procedure for FDR control. Since the null hypothesis enables computation of estimates using the quadratic programming framework, we use both the nonparametric (residual) bootstrap and parametric bootstrap (using t-distributed noise) to compute two versions of p-values. We found that that the rankings of p-values by the two different variants of bootstrap procedures were highly positively correlated, which means our method is robust and not sensitive to the way we compute p-values.

All indicated significant pairs by the residual bootstrap and percentile bootstrap CI method are shown in Tables 5.3 and 5.4. The results related to using the null bootstrap distribution to compute p-values are shown in the Supplementary Materials (Tables A.1 and A.2). It is interesting to see that two versions of p-values indicate similar results from residual bootstrap and parametric bootstrap, respectively. However, the nonparametric (residual) bootstrap leads to a more conservative conclusion (seven to eight significant pairs), compared to the parametric one (13 significant pairs).

Simultaneous Inference—Maximum "Unweighted" Productivity ψ_i

We now demonstrate that a wider variety of applications can be tackled by the proposed framework. For example, we can make inferences related to the maximum "unweighted" productivity. We consider using the maximum "unweighted" productivity

$$\psi_i := \max_t h_i^u(t) \tag{5.13}$$

	p-value by percentile bootstrap CI	CI lower	CI upper	CI length	mean	sd	est
$\hat{\tau}_A^{bs} - \hat{\tau}_H^{bs}$	$<10^{-5}$	−28.059	−10.204	17.856	−19.298	3.624	−19.186
$\hat{\tau}_B^{bs} - \hat{\tau}_H^{bs}$	$<10^{-5}$	−25.265	−8.512	16.753	−17.707	3.213	−18.115
$\hat{\tau}_C^{bs} - \hat{\tau}_H^{bs}$	$<10^{-5}$	−26.081	−9.485	16.596	−18.197	3.379	−18.347
$\hat{\tau}_D^{bs} - \hat{\tau}_F^{bs}$	0.008	0.659	17.297	16.638	8.125	3.348	7.544
$\hat{\tau}_D^{bs} - \hat{\tau}_H^{bs}$	$<10^{-5}$	−19.830	−3.745	16.086	−12.414	3.191	−12.679
$\hat{\tau}_E^{bs} - \hat{\tau}_H^{bs}$	$<10^{-5}$	−24.668	−7.003	17.665	−15.496	3.489	−15.233
$\hat{\tau}_F^{bs} - \hat{\tau}_H^{bs}$	$<10^{-5}$	−29.730	−11.710	18.020	−20.539	3.677	−20.223
$\hat{\tau}_G^{bs} - \hat{\tau}_H^{bs}$	$<10^{-5}$	−24.095	−7.635	16.460	−16.482	3.187	−16.687

Table 5.3. Using **residual** (nonparametric) bootstrap methods: False coverage-statement rate (FCR)—Adjusted BH-Selected CIs for selected parameters indicated by **the percentile bootstrap CI**; All confidence intervals above show significance against $H_0 : \tau_i = \tau_j$.

	p-value by percentile bootstrap CI	CI lower	CI upper	CI length	mean	sd	est
$\hat{\tau}_A^{bs} - \hat{\tau}_D^{bs}$	0.002	−10.778	−3.105	7.674	−6.578	1.545	−6.507
$\hat{\tau}_A^{bs} - \hat{\tau}_H^{bs}$	$<10^{-5}$	−23.481	−15.508	7.974	−19.277	1.731	−19.186
$\hat{\tau}_B^{bs} - \hat{\tau}_D^{bs}$	0.006	−9.552	−1.525	8.026	−5.301	1.525	−5.436
$\hat{\tau}_B^{bs} - \hat{\tau}_H^{bs}$	$<10^{-5}$	−22.067	−14.185	7.882	−17.999	1.653	−18.115
$\hat{\tau}_C^{bs} - \hat{\tau}_D^{bs}$	0.006	−10.075	−1.356	8.719	−5.662	1.644	−5.668
$\hat{\tau}_C^{bs} - \hat{\tau}_H^{bs}$	$<10^{-5}$	−22.761	−14.358	8.403	−18.361	1.709	−18.347
$\hat{\tau}_D^{bs} - \hat{\tau}_F^{bs}$	0.002	2.768	12.825	10.057	7.680	1.828	7.544
$\hat{\tau}_D^{bs} - \hat{\tau}_G^{bs}$	0.022	0.383	8.532	8.149	4.106	1.553	4.008
$\hat{\tau}_D^{bs} - \hat{\tau}_H^{bs}$	0.002	−16.687	−8.478	8.209	−12.698	1.765	−12.679
$\hat{\tau}_E^{bs} - \hat{\tau}_F^{bs}$	0.018	0.121	10.278	10.157	4.859	1.877	4.990
$\hat{\tau}_E^{bs} - \hat{\tau}_H^{bs}$	$<10^{-5}$	−20.403	−11.297	9.106	−15.519	1.925	−15.233
$\hat{\tau}_F^{bs} - \hat{\tau}_H^{bs}$	$<10^{-5}$	−25.471	−15.836	9.635	−20.378	1.872	−20.223
$\hat{\tau}_G^{bs} - \hat{\tau}_H^{bs}$	$<10^{-5}$	−20.893	−12.667	8.225	−16.804	1.648	−16.687

Table 5.4. Using **parametric** bootstrap: False coverage-statement rate (FCR)—Adjusted BH-Selected CIs for selected parameters indicated by **the percentile bootstrap CI**; All confidence intervals above show significance against $H_0 : \tau_i = \tau_j$.

as the quantity of interest, where

$$h_i^u(t) := \frac{\mu_i(t)}{(t + T_i^{(c)})} \tag{5.14}$$

$T_i^{(c)}$ is the days of cultivation before the induction.

The estimated "unweighted" productivity curves are shown in Figure 5.6. In this case, since the parameter is a nonlinear function of the trajectories, it is not possible to obtain the p-values using the null bootstrap distribution option. Instead, we use the correspondence between hypothesis testing and finding confidence intervals.

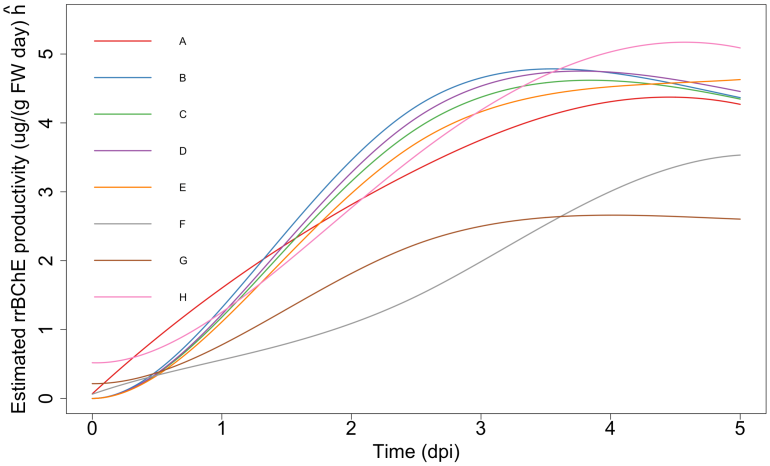


FIGURE 5.6. Estimated "unweighted" productivity curves based on observed data. Note that $T^{(c)} = (6, 6, 6, 7, 6, 8, 14, 8)$.

We apply the same procedures that we used for the maximum production level to the maximum "unweighted" productivity, using both the parametric bootstrap and residual bootstrap methods. Tables A.3 and A.4 (in the Supplementary Material) show that residual bootstrap method leads to more conservative results (14 significant pairs) compared to the parametric one (21 significant pairs).

5.3.3. Summary of Findings. We summarize the findings from the real data analysis as follows:

Comparison between Two Versions of p-Values

The results of the two bootstrap methods imply that the rankings of two versions of p-values are highly positively correlated ($\geq$0.96), which means that our method is robust and not sensitive to the way in which we compute p-values.

Comparison of Residual Bootstrap and Parametric Bootstrap

Tables A.1 and A.2 may be compared for a performance of different bootstrap strategies. First, within each table, we can see the p-values by using the null and duality of p-value and CI are consistent based on the rank and its correlation. Second, we can see that residual bootstrap is more conservative (Table A.1 only has seven significant differences), and the p-values are larger. We can see the main difference depends on the variability of data (the way we generate bootstrap samples). Parametric bootstrap samples show higher variability (we make use of the distribution assumption), while residual ones have lower variability. This means that the residual bootstrap is more conservative, and the result is consistent with Figure 5.7.

Compared to the parametric bootstrap method, the residual bootstrap method is more conservative (see Tables A.1 and A.2). Comparing the parametric and residual bootstrap sampling distributions for parameters (Figure 5.7), it is clear that the residual bootstrap method leads to more widely spread sampling distributions, which yields larger p-values and fewer significant testing results. Again, the residual bootstrap method is not particularly effective when the number of factors is small. $I = 8$ might not be sufficient, and a choice needs to be made to make inferences.

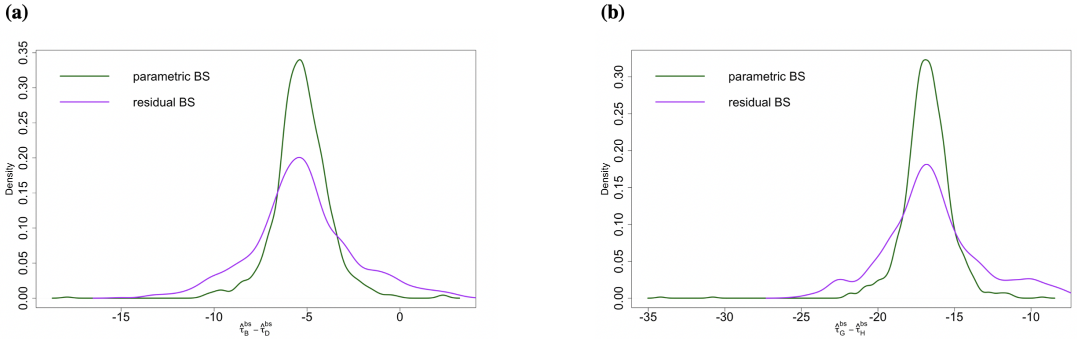


FIGURE 5.7. Comparison: **empirical** parametric bootstrap sampling distribution and **empirical** residual bootstrap sampling distribution of (**a**) $\hat{\tau}_B^{bs} - \hat{\tau}_D^{bs}$, which is significant in the parametric bootstrap case but not in the residual bootstrap case. (**b**) $\hat{\tau}_G^{bs} - \hat{\tau}_H^{bs}$, which is significant in both cases.

5.4. Discussion

Implications for scientific investigation: The analysis presented here shows the merit of using advanced statistical techniques for answering critical questions about the comparative effectiveness of different experimental constructs in complex and expensive biological experiments where the data availability is limited. In particular, our study shows the capability of well-designed bootstrap methods for obtaining accurate confidence intervals for parameters of interest. In addition, it shows that the shapes of the mean protein production trajectories have a direct influence on the variability of the estimates of the "optimal time-to-harvest" and related parameters.

Moreover, the simulation study shows that the specialized resampling procedures provide a good description of this variability and therefore help experimenters to formulate an appropriate experimental design in terms of the number of time points and replicates for the experiment to achieve a desired level of accuracy. These assertions are further validated by an application of the methods to analyze the rrBChE data.

Furthermore, a salient feature of our framework is that the methodology works well even when the data are limited, and thus the standard large sample theory for statistical inference is not applicable, which is common in complicated biological experiments. In particular, the proposed methodology can help to suggest and validate the strategy of designing lengthy and expensive experiments where the number of replicates is limited. We also found that the two variants of bootstrap methods work well even when the amounts of data are limited. Specifically, if one is more confident in the model assumption, then parametric bootstrap is the preferred option. Otherwise, the residual bootstrap is a better choice as it yields more conservative confidence intervals.

In addition, our model can be used to obtain the optimal set of conditions to maximize parameters of interest, such as the optimum harvest time. When conditions are considered discrete, our resulting model framework has capabilities to provide statistical simultaneous inference to make comparisons. When the conditions are continuous, and the rate of change in the optimum harvest time is to be measured, we can modify the model by treating the optimum harvest time as a function of these conditions (treated as continuous covariates in the model).

Then, to optimize the optimum harvest time, we look to solve an equation setting the derivative (with respect to the conditions) of the optimum harvest time to zero. In that case, we can still apply an appropriately modified form of our bootstrap procedure to obtain confidence intervals for the parameters of interest. This can be also validated experimentally if enough resources are allowed in practice.

Finally, we discuss two possible directions in which the methodology presented here can be further enhanced to deal with additional questions about data generated from similar biological experiments.

Extension 1: The preceding analysis makes it clear that our framework is highly effective particularly when multiple process metrics are of interest. In addition to the parameters we discussed, our framework can be extended to other parameters, e.g., the average production and average "unweighted" productivity. The average production for the i-th factor is described as $\int_{T_0}^{T_{\max}} \mu_i(t)dt$ where T_0 is the starting point.

The corresponding null hypothesis is $\int_{T_0}^{T_{\max}} \mu_1(t)dt = \cdots = \int_{T_0}^{T_{\max}} \mu_I(t)dt$. Similarly, the average "unweighted" productivity is given by $\int_{T_0}^{T_{\max}} \mu_1(t)/(t + T_i^{(c)})dt$ for the i-th factor. The null hypothesis is: $\int_{T_0}^{T_{\max}} \mu_1(t)/(t+T_i^{(c)})dt = \cdots = \int_{T_0}^{T_{\max}} \mu_I(t)/(t+T_i^{(c)})dt$. However, p-values cannot be calculated based on the null bootstrap distribution for average "unweighted" productivity because of nonlinearity. Instead, we may adopt the percentile bootstrap option to obtain p-values.

Extension 2: As we mentioned earlier, we used a simplified version of ANOVA framework. This can be extended to a two-factor ANOVA framework by incorporating the effect of the experimental condition, time and their interactions, with appropriate constraints. Such a modified model is able to leverage the impact of different factors and interactions of both factor and time and how different they are across factors. These are also key interests of experimental studies. In addition to Equations (5.1)–(5.3), we incorporate the factor effects and factor–time interaction by modeling α_{il} as follows:

$$\alpha_{il} = \beta + \eta_i + \xi_l + \delta_{il}, \qquad i = 1, \dots, I; l = 1, \dots, L \tag{5.15}$$

To ensure identifiability of the parameters, we impose the following linear constraints:

$$\begin{aligned} &\sum_{i=1}^{I} \eta_i = 0, \quad \sum_{l=1}^{L} \xi_l B_l(t^*) = 0 \\ &\sum_{i=1}^{I} \delta_{il} = 0 \quad \text{for each} \quad l, \quad \sum_{l=1}^{L} \delta_{il} B_l(t^*) = 0, \quad \text{for each} \quad i, \end{aligned}$$

where t^* is an arbitrary but appropriately chosen point in $[0, T]$. Furthermore, the constraints (5.3) are equivalent to

$$\xi_{l+1} - \xi_l + \delta_{i(l+1)} - \delta_{il} \geq 0, \qquad i = 1, \dots, I; \ l = 1, \dots, L-1. \tag{5.16}$$

The extended model given by (5.1), (5.2), (5.15) and (5.16) is a modified form of two-factor ANOVA and can be solved by quadratic programming. Statistical inference can be conducted by a natural extension of the resampling strategy proposed here.

CHAPTER 6

Extension of Analysis of variability of functionals of recombinant protein production trajectories based on limited data

6.1. Introduction

Biological experiments that involve the production of a certain recombinant molecule over a period of time under different experimental conditions are common in research. Hence, the data that results from such experiments are inherently longitudinal, meaning that they are recorded over a period of time. One key challenge in analyzing such data is to compare the optimum trajectories across different factors or experimental conditions or treatments, which is a fundamental topic of longitudinal data analysis [**38**, **48**, **65**, **73**]. Typically, the objective is to estimate the expected amount of the ingredient being measured, and multiple replicates are necessary to accommodate a comprehensive Analysis of Variance (ANOVA) approach to solve the problem to account for the effects of various factors. Traditionally, repeated measures designs have been the common approach to answer such inferential questions [**42**, **93**, **105**]. Nevertheless, in many real-life lab-based biological experiments, there is often a limitation in the number of data points or replicates that can be obtained. This constraint is due to the time, costs, and resources associated with completing each condition or treatment, particularly in scaling-up experiments. Moreover, the key object of interest can be either the target molecule itself but could also be functions of the trajectories. As such, alternative approaches that allow for fewer data points or replicates while still producing accurate results are needed. This is particularly important as researchers seek to study and understand larger and more complex systems, where obtaining a large number of data points or replicates is not always feasible.

The hydrolase enzyme known as butyrylcholinesterase (BChE) is present in human blood plasma as a tetramer and has the potential to be used as a prophylactic or therapeutic measure against organophosphorus nerve agent poisoning [**74**]. However, the use of purified BChE obtained from human blood plasma in clinical settings is restricted by its high cost, estimated to be $20,000 per 400 mg dose [**3**]. To overcome this limitation, researchers have developed recombinant human BChE (rBChE) using various host expression systems, such as transgenic rice cell suspension cultures, as an alternative source of BChE. A rice-made recombinant human BChE (rrBChE) has been developed in the lab [**19**,**76**]. However, the experiments are expensive and time-consuming because of the slow growth rate of plant cells in the lab environment. As a result, the number of replicates or units is restricted and thus it can be challenging to interpret the data and integrate the information in practice. To understand the data in a better way, it is crucial to develop a comprehensive statistical framework to characterize the production curves over time, where the time points or replicates are limited practically.

Hence, we develop a flexible simplified ANOVA framework that is effective for limited data or replicates to model the mean trajectories as well as make inferences on parameters [**72**]. It contributes effectively to employing statistical methods to address vital inquiries regarding the relative efficacy of various experimental designs in elaborate and costly biological experiments, which are often restricted by limited data availability. Specifically, the efficacy of well-crafted bootstrap techniques in generating precise confidence intervals for the parameters of interest is demonstrated. Moreover, this approach reveals that the forms of the average production trajectories have a significant impact on the precision of the estimates of key parameters such as "optimal time-to-harvest" and others that are closely related. Furthermore, this model has the potential to be utilized for determining the ideal experimental conditions that can optimize key parameters of interest. In situations where the conditions are distinct and can be grouped, this simplified model framework can provide statistical inference simultaneously to facilitate meaningful comparisons.

However, the aforementioned ANOVA framework is a simplified version. It fails to answer some other questions that are also key studies of interest. For example, researchers may be interested in quantifying the treatment and time effect. In addition, since the trajectory is affected by treatment and time as two factors, one may naturally raise a question of the presence of interaction between treatment and time effect. It is necessary to investigate how different treatment effects are. The simplified version is not capable to answer these questions because the coefficients represent a composite effect of treatment and time effect. Notably, it is crucial to consider the dependence in terms of time, while the current study so far only involves independent observations in the simplified version. In practice, it is reasonable to consider the observations possibly correlated over time as this is intrinsically longitudinal repeated-measurement data [**29**,**59**,**60**]. It is common to account for the temporal time series over the experiments. In addition to the regular covariates, autocorrelation also plays a role in determining the true shape of the trajectory and explaining the variation from different sources. In such biological experimental settings, correlated longitudinal data usually comes with the characteristics of decaying autocorrelation over time. A lot of correlation structures have been proposed and applied to repeated-measure designs. Examples of common structures of correlation are AR(1), damped exponential (DE) structure and other structures in Gaussian repeated measure context [**24**, **54**, **75**, **88**, **96**]. All correlation structures have their own merits for characterizing the variabilities. Although a specific correlation pattern may be suitable for certain research questions or study designs, it may not be effective for others. We need to be cautious to select one that closely connects to the scientific field and the context of our study design.

Therefore, a comprehensive framework is essential and necessary for answering the questions, given the limited data for such biological experiments. Thanks to the flexibility of the simplified structure, we are able to easily extend it to a more generalized statistical framework by incorporating the two-factor ANOVA framework [**34**, **97**], where the treatment effect, time effect and their interaction are modeled. The observations can be either independent or correlated over time by specifying the covariance matrix in the model, leading to a possibly weighted quadratic programming problem with still linear constraints for shape requirements. We consider a family of AR(1) covariance structures with 2 or 3 variations coefficients in that free structures may result in over-parameterization issues with limited data. This family structure is known as most useful in many longitudinal studies to quantify the variation. Besides, it allows for two types of various sources that are practically present in biological experiments. Given this flexible structure, we deploy a two-step estimation where an optimization algorithm is used for the maximum likelihood (ML) estimation of covariance coefficients and a weighted quadratic programming algorithm is implemented for regular coefficients. Under the limited data settings, the inference can be conducted in similar ways as for the simplified version. A bootstrap-based F-test is constructed for testing the presence of the interaction. However, it is imperative for the choice of distribution to assume in parametric bootstrap when it comes to correlated structures. A wide spread of the variance is shown in our numerical study and thus using Gaussian context may fail to capture the whole variability. We show that t-distribution is a better option for this case. To achieve this, we develop a bagging-based strategy to select the optimal number of degrees of freedom, accordingly.

The rest of the content is organized as follows: we review the simplified ANOVA framework and discuss the generalized two-factor ANOVA model and inference procedures in Section 6.2. In Section 6.3, we illustrate the modeling and inference on the questions of interest through numerical studies.

6.2. Statistical Methodology

We suppose that there are $I \geq 2$ factors (or, treatments) and each treatment is applied to several independently chosen experimental units (bioreactors). Further, the response (e.g. rrBChE concentration in the bioreactor) is measured at observation times $0 < t_{i1} < \cdots < t_{iJ} = T$, say, for $J \geq 2$ (this allows the observation times to be different for different factors). Let us denote the mean response curve (at time $t \geq 0$) corresponding to the i-th factor as $\mu_i(\cdot)$. It is assumed that $\mu_i(\cdot)$ is a monotonically increasing function of time, over the observation time window $[0, T]$.

For simplicity as well as statistical efficiency, we assume a *balanced experimental design*, that is, the sample size at each observational time is the same for every treatment. The measurement process is destructive, so that for any particular experimental unit, we only have one measurement, at the time of sampling the bioreactor. Therefore, to obtain reasonably accurate measurement for the whole trajectory, we need replicates (i.e., multiple experimental units), for each time t_{ij} and each treatment i. Let n denote the number of replicates assigned to each combination (i, j), which corresponds to a balanced design. Note that we allow $n = 1$ since in practice only limited data are available especially in some biological experiments. Denote the response from the k-th experimental unit, in the i-th factor group, measured at time t_{ij} , to be Y_{ijk}.

Suppose each treatment is applied to several independently chosen experimental units. The response under the i-th treatment is measured at observation time $0 < t_{i1} < \cdots < t_{ij} < \cdots < t_{iJ} = T$, for $J \geq 2$.

6.2.1. Review: Simplified ANOVA Framework with Monotonicity Constraints. With the goal of estimations and inference of the varying functional, such as "optimal time-to-harvest", we proposed a comprehensive ANOVA-based statistical framework to model the production curves, where linear constraints are applied due to the characteristics of the growing responses. The model framework is as follow:

$$Y_{ijk} = \mu_i(t_{ij}) + \epsilon_{ijk}, \quad j = 1, \ldots, J; k = 1, \ldots, n; i = 1, \ldots, I. \tag{6.1}$$

$$\mu_i(t) = \Sigma_{l=1}^{L} \alpha_{il} B_l(t), \quad i = 1, \ldots, I \tag{6.2}$$

$$\alpha_{i(l+1)} \geq \alpha_{il} \quad \text{for} \quad l = 1, \ldots, L-1 \tag{6.3}$$

For statistical inference on the parameters of interest, we use *bootstrap*, or resampling procedures. We compare the efficacies of several different versions of bootstrap, namely, the residual bootstrap, parametric bootstrap and nonparametric bootstrap. These methods are well-adapted to deal with the problem of comparing the values of parameters that are arbitrary functionals of the production trajectories corresponding to different factors. Moreover, we provide a comprehensive framework for simultaneous statistical inference on several process metrics that are functionals of biochemical production (or growth) trajectories, based on fairly limited amount of data, with empirical validity. More details of the method and numerical study can be found in Chapter 5.

6.2.2. General Two-factor ANOVA Framework with Monotonicity Constraints. Note that the aforementioned model is a simplified version of a more general ANOVA framework that enables quantifying possible interactions between the treatments and time, with appropriate constraints. In addition to address the mentioned parameters of interest, which are the functionals of production curves, such a modified model is able to leverage the impact of different factors and interactions of both factor and time and how different they are across factors. These are also key interests of experimental studies.

We derive the modified version of two-factor ANOVA model by incorporating treatment effects and time factor. This is an extension of one-factor ANOVA with monotonic trajectories studied by Chapter 5. The model are specified as followed:

In addition to equation (6.1), (6.2) and (6.3), We incorporate the factor effects and factor-time interaction by modeling α_{il}'s as follows:

$$\alpha_{il} = \beta + \eta_i + \xi_l + \delta_{il}, \qquad i = 1, \ldots, I; l = 1, \ldots, L \tag{6.4}$$

To ensure identifiability of the parameters, we impose the following linear constraints:

$$\begin{aligned} &\sum_{i=1}^{I} \eta_i = 0, \quad \sum_{l=1}^{L} \xi_l B_l(t^*) = 0 \\ &\sum_{i=1}^{I} \delta_{il} = 0 \quad \text{for each} \quad l, \quad \sum_{l=1}^{L} \delta_{il} B_l(t^*) = 0, \quad \text{for each} \quad i, \end{aligned}$$

where t^* is an arbitrary but appropriately chosen point in $[0, T]$. And the constraints (6.3) are equivalent to

$$\xi_{l+1} - \xi_l + \delta_{i(l+1)} - \delta_{il} \geq 0, \qquad i = 1, \ldots, I; \ l = 1, \ldots, L-1. \tag{6.5}$$

The extended model given by (6.1), (6.2), (6.4) and (6.5) is a modified form of two-factor ANOVA and can be solved by quadratic programming. Statistical inference can be carried out by a natural extension of the resampling strategy proposed in 5.

6.2.3. Correlated observations across experimental time. The observations from the experiments measured over time practically can be correlated. Thus introducing a correlated structure is more effective to capture the correlated data pattern. The structure of the variation thus consists of two sources of noise: one is measurement noise while the other is temporal noise. We extend our standard framework by applying a correlated structure, which can be modeled by introducing variance and correlation coefficients. In addition to the usual effect parameters (e.g., treatment effects) we are interested in, we are able to estimate variance and correlation coefficients and conduct the corresponding inference for them with *bootstrap* procedures.

A general covariance structure Suppose the data within the $i-$th treatment, Y_i's are correlated in terms of time. With the current model given by (6.1), (6.2), (6.4) and (6.5), it can be modified as follows:

$$Y_i = (Y_{i11}, Y_{i12}, ..., Y_{iNJ})^T, \epsilon_i = (\epsilon_{i11}, \epsilon_{i12}, ..., \epsilon_{iNJ})^T \tag{6.6}$$

where $\epsilon_i \sim N(0, \Sigma_i)$.

The elements of Σ_i are in a form of three variance parameters:

$$cov(\epsilon_{ikj}, \epsilon_{ikj'}) = \sigma_\epsilon^2 \tilde{\delta}_{jj'} + \sigma_a^2 \rho_\theta^{|t_{ikj} - t_{ikj'}|} \tag{6.7}$$

where $\tilde{\delta}_{jj'} = 1$ if $j = j'$, 0 otherwise, for all $1 \leq i \leq I, 1 \leq k \leq N, 1 \leq j \leq J$. The proposed structure in Eq(6.7) indicates two sources of variation: σ_ϵ^2 is to explain the systematic error while σ_a^2 quantifies the measurement error.

Two-step estimation The estimation of both variance parameters and regular factor coefficients can be achieved by using a two-step estimation:

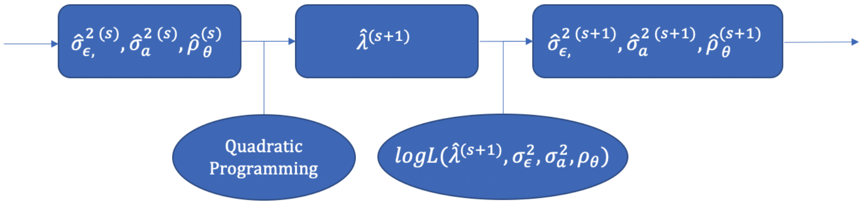

$\hat{\sigma}_{\epsilon,}^{2\ (s)}, \hat{\sigma}_a^{2\ (s)}, \hat{\rho}_\theta^{(s)}$
$\hat{\lambda}^{(s+1)}$
$\hat{\sigma}_{\epsilon,}^{2\ (s+1)}, \hat{\sigma}_a^{2\ (s+1)}, \hat{\rho}_\theta^{(s+1)}$
Quadratic Programming
$logL(\hat{\lambda}^{(s+1)}, \sigma_\epsilon^2, \sigma_a^2, \rho_\theta)$

FIGURE 6.1. A flow chart indicating the two-step estimation for correlated structure.

A simplified covariance structure Though Eq(6.7) describes a relative comprehensive variation, the variance parameters can be hard to estimate due to the complicated model structure and limited data in realistic. A compromise can be made by simplifying the Eq(6.7) by dropping the system variance

$$cov(\epsilon_{ikj}, \epsilon_{ikj'}) = \sigma_a^2 \rho_\theta^{|t_{ikj} - t_{ikj'}|} \tag{6.8}$$

And thus the algorithem can be simplified as shown in Fig 6.2

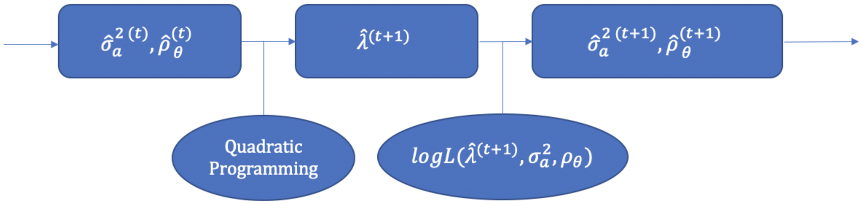

$\hat{\sigma}_a^{2\,(t)}, \hat{\rho}_\theta^{(t)}$
$\hat{\lambda}^{(t+1)}$
$\hat{\sigma}_a^{2\,(t+1)}, \hat{\rho}_\theta^{(t+1)}$
Quadratic Programming
$logL(\hat{\lambda}^{(t+1)}, \sigma_a^2, \rho_\theta)$

FIGURE 6.2. A flow chart indicating the simplified two-step estimation for correlated structure.

6.2.4. Bootstrap procedures. We adopt parametric, residual bootstrap procedures for inference, given the practical data size. Biochemical experiments are often time consuming and expensive and thus the number of experimental units is usually extremely limited, where central limit theory is not applicable. The detailed steps for the simplified ANOVA framework in Chapter 5 can be easily extended for the general two-factor ANOVA framework. With the bootstrap samples, we are able to conduct the testings of interest via bootstrap distributions, bootstrap $p-$values as well as bootstrap confidence intervals (CIs).

bootstrap with the assumption of t-distributed responses In addition to assume a Gaussian structure of the distribution of the responses, we also consider multivariate $t-$distribution to allow for more extreme values. This can be realized by generating $t-$distributed noise, instead of Gaussian noise, for bootstrap samples. However, the number of degrees of freedom needs to be determined to generate bootstrap samples and it is challenging to estimate an appropriate number of degrees of freedom through an intrinsic modeling. To achieve this, we split the whole data set into a training set and test set and thus we will apply the following steps for each candidate of possible degree of freedoms. With one possible number of degrees of freedom, the resulted estimations based on the training set will be applied to the test set for generating the bootstrap prediction intervals of each observations. We will choose the optimal number of degrees of freedom obeying the following rules:

- Suppose $\{\nu_1, \ldots, \nu_D\}$ as the candidates of the number of degrees of freedom, and the selection threshold of the coverage rate is $\tilde{C}$,
- Leave N_{test} data points out of the full data and consider them as a testing set while the remaining as the training set
- Apply the regular procedures to the training set

- Use the estimates obtained from the previous step in t-bootstrap procedures with $\nu_d \in \{\nu_1, \ldots, \nu_D\}$
- Predict the testing set based on each bootstrap estimates and thus obtain $(1-\alpha)100\%$ prediction intervals for each sample of the testing set, where α is a pre-determined significance level.
- Compute the according coverage rate, $\frac{n_c(\nu_d)}{N_{test}}$ where $n_c(\nu_d)$ denotes the number of prediction intervals covering the observations of the test set.
- Repeat the steps above for all $\nu_d \in \{\nu_1, \ldots, \nu_D\}$
- Select the optimal number of degrees of freedom $\hat{\nu}_d = \max\{\nu_d : \frac{n_c(\nu_d)}{N_{test}} \geq \tilde{C}\}$

We repeat the steps above for multiple times where each time an optimal number of degree of freedom is suggested. We will select the one that has been recommended the most. This "bagging" approach helps stabilize the selection.

Due to the characteristics of t-distribution, it is known that a t-distribution converges to the normal distribution, when the number of degrees of freedom increases. Therefore, the larger number of degrees of freedom tends to yield narrower prediction intervals. This will lead to a relative conservative choice of the number of degrees of freedom when we hope to maintain a good coverage rate of the testing data. Though the choice could be biased, we demonstrate the impact is negligible later in the numerical study.

6.2.5. Inference. Similar to the simplified framework, the current comprehensive framework is able to make similar inference for the interested functionals of production curves such as "optimal time-to-harvest" and "maximum production". Moreover, it has capability to test the pairwise comparisons of treatment effect and presence of interaction effect, which are not available in simplified version. This allows us to investigate more about the underlying characteristics of the experiments, such as the differences among treatment factors and the impact resulted by the treatment or time factor.

Simultaneous inference for the treatment factor The hypothesis of pairwise comparisons of treatment effect is constructed as follows:

$$H_0 : \eta_i = \eta_{i'} \text{ for any } i \neq i' \text{ vs } H_a : \text{at least one } \eta_i \text{ is different}$$

We proceed the testing with (parametric or residual) bootstrap procedures. With multiple tests, it becomes necessary to adjust the level of significance of each individual test. This can be achieved by making use of the *Bonferroni procedure* or a *False Discovery Rate (FDR) control* procedure [**28**]. Once we obtain the bootstrap p-values for the different tests, the *Benjamini-Hochberg (BH) procedure* for FDR control is used to first determine the significance level for each pairwise test for a given level of familywise significance. Thus, the confidence levels of the confidence intervals (*False Coverage-statement Rate* (FCR)-Adjusted BH-Selected CIs) for the parameters are adjusted accordingly [**10**]. The desciption of the procedures are available in [**72**].

Test the presence of interactions The hypothesis can be described as

$$H_0 : \delta_{ij} = 0 \text{ for all } 1 \leq i \leq I, 1 \leq j \leq J \text{ vs } H_a : \text{at least one } \delta_{ij} \neq 0$$

The common test to be used is F-test for $H_0 : \delta_{11} = \cdots = \delta_{IJ} = 0$. It indicates the degree of the difference between the reduced model, where H_0 is assumed to be true, and the full model with at least one $\delta_{ij} \neq 0$. Due to the limited size, the theoretical F-statistics and its approximate distribution are not valid anymore and therefore we adopt a bootstrap-based strategy to conduct F-test. The idea is to compare the observed F-statistics to the bootstrap-based F-distribution. The detailed steps are described below for the computation:

- Obtain the fitted curves $\hat{\mu}_i(t)$ based on the original data with the regular proposed procedures;
- Obtain the fitted curves $\tilde{\mu}_i(t)$ based on the original data by imposing additional constraints under $H_0 : \delta_{ij} = 0$;
- Generate the surrogate non-parametric bootstrap data for multiple times (e.g., $B^s = 1000$) $\tilde{Y}^b_{ijk} = e^b_{ijk} + \tilde{\mu}_i(t_j)$ where e^b_{ijk} are sampled from $\{e_{ijk} - \tilde{e}_j\}$ and $e_{ijk} = Y_{ijk} - \hat{\mu}_i(t_j)$, $\tilde{e}_j = \frac{1}{I*N} \sum_i \sum_k e_{ijk}$. Note that $\tilde{Y}^b_{ijk}$'s are considered the bootstrap samples under H_0;
- Compute the observed $F-$statistics, F^{obs}, and the null distribution consisting of the $F-$statistics depending on B^s surrogate non-parametric bootstrap data, $\{F^s_b\}^{B^s}_{b=1}$;
- Obtain the $p-$value $= \frac{\#\{F^s_b > F^{obs}\}}{B^s}$

6.3. Numerical experiments

In this section, we display the simulated data and according results. We conduct tests for normal and t-distributed data, respectively. All the results indicate the consistency of our model results with the experimental settings, demonstrating the capability of the model framework, in estimation and inference, with limited conditions.

6.3.1. Independent observations with Gaussian assumption. We generate a set of data where normal assumption and independence are true. We fit the data with our model framework. The fitting is shown in Figure 6.3. Next, we report the simultaneous testing results of $H_0 : \eta_i = \eta_{i'}$. The fitted curves obtained are close to the true ones, indicating the effectiveness in production fitting.

Given the situation of limited data, the large number theory is not reliable for inferential questions. Instead, we turn to bootstrap approaches to generate a relatively large amount of samples that inherit similar characteristics as the observations, and then conduct hypothesis testing of interest. As we described before, two main types of bootstrap procedures are used: parametric bootstrap and residual bootstrap. The former is implemented by generating bootstrap samples under an assumed distribution (usually a Gaussian or $t-$ distribution), with estimated distribution parameters determined by fitting the observations. In comparison, the residual bootstrap is non-parametric, and we generate bootstrap samples by repeatedly sampling residuals obtained from the fitted model. The differences between the two bootstrap approaches are minor, though the residual bootstrap may yield more conservative results (e.g., larger variations). We demonstrate the robustness and capability of both bootstrap procedures in characterizing the true distributions in Figure 6.4.

We conduct both parametric and residual bootstrap procedures for the simultaneous testing in the treatment effect, and it is shown both procedures imply stable and valid results. The BH procedure for FDR control is applied and the adjusted confidence intervals for the treatment difference are computed based on two types of method for p-value calculation, which is shown in Table 6.1, Table 6.2, Table 6.3, Table 6.4. Though there are some minor differences in p-values and adjusted confidence intervals, the conclusions are the same in terms of the hypothesis. We

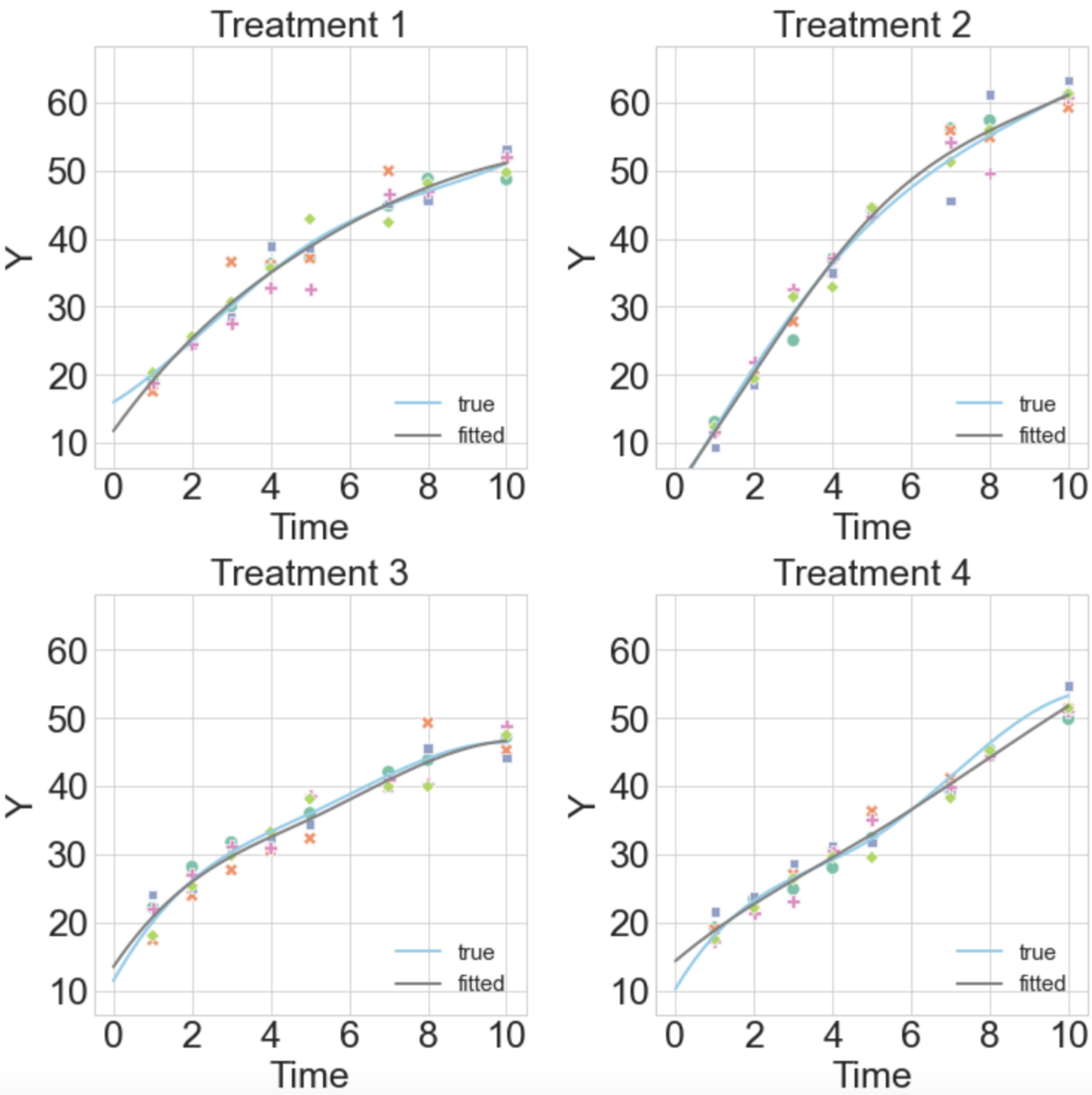


FIGURE 6.3. Observations marked as dots across $I = 4$ treatments and $J = 8$ time points. The fitted curves are in black, compared to the true curves in blue.

reject all hypotheses $H_0 : \eta_i = \eta_{i'}$ except for $H_0 : \eta_3 = \eta_4$, indicating their similar effect on the production. It is consistent to the true and reasonable as the true difference of these two treatment effects is 0.2 which is so small that it is hard to differentiate them. The results also indicate that the choice of bootstrap procedures for the current model framework does not impact the conclusions significantly.

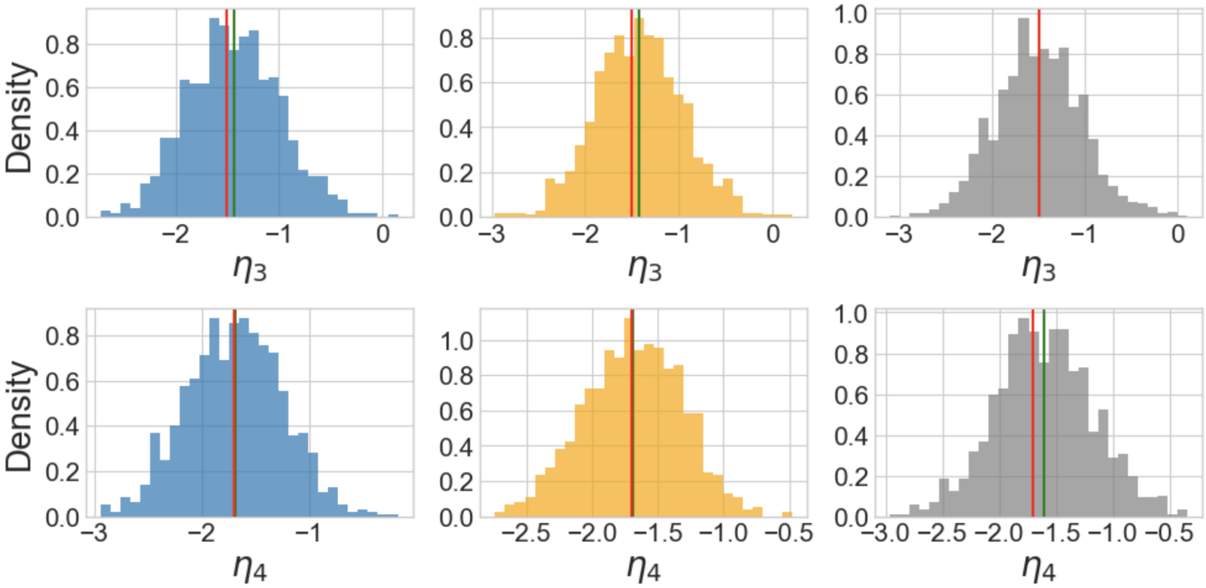


FIGURE 6.4. Bootstrap distribution by the parametric bootstrap in blue; bootstrap distribution by the residual bootstrap in yellow, and the true sampling distributions in grey of η_3, η_4 (green vertical line: the sample mean; red vertical line: the true parameter).

	TRUE	mean	std	est	p-value	CI lower	CI upper	CI length	reject?
$\eta_1 - \eta_2$	-1.2	-2.0383	0.7696	-2.0430	0.007	-3.549	-0.402	3.147	Yes
$\eta_1 - \eta_3$	2.5	1.9758	0.7373	2.0172	0.003	0.457	3.393	2.936	Yes
$\eta_1 - \eta_4$	2.7	2.2372	0.7500	2.2547	0.001	0.695	3.825	3.130	Yes
$\eta_2 - \eta_3$	3.7	4.0141	0.7400	4.0602	$< 10^{-5}$	2.498	5.440	2.942	Yes
$\eta_2 - \eta_4$	3.9	4.2756	0.7255	4.2977	$< 10^{-5}$	2.776	5.802	3.026	Yes
$\eta_3 - \eta_4$	0.2	0.2615	0.7381	0.2375	0.721	-1.187	1.738	2.925	No

TABLE 6.1. Using **parametric** bootstrap: False coverage-statement rate (FCR) - Adjusted BH-Selected CIs for selected parameters indicated by **the percentile bootstrap CI**; All confidence intervals above show significance against $H_0 : \eta_i = \eta_{i'}$

In addition to the testing above, we use the proposed bootstrap-based F-test to conduct the hypothesis testing for the presence of interaction $H0 : \delta_{il} = 0$. We display two simulation results under different settings: one is generated from the model with all interactions exist, while the other is generated from the model with 0 interactions. For both settings, we run the testing and compute the observed F-statistics and bootstrapped ones under the null hypothesis, where the latter form the null bootstrap-based F-distribution. The p-values are computed accordingly. The

	TRUE	mean	std	est	p-value	CI lower	CI upper	CI length	reject?
$\eta_1 - \eta_2$	-1.2	-2.0445	0.6610	-2.0430	0.000	-3.359	-0.652	2.707	Yes
$\eta_1 - \eta_3$	2.5	1.9618	0.7426	2.0172	0.009	0.320	3.464	3.144	Yes
$\eta_1 - \eta_4$	2.7	2.2259	0.6320	2.2547	0.000	0.939	3.481	2.542	Yes
$\eta_2 - \eta_3$	3.7	4.0063	0.7363	4.0602	$< 10^{-5}$	2.481	5.403	2.922	Yes
$\eta_2 - \eta_4$	3.9	4.2704	0.6017	4.2977	$< 10^{-5}$	3.011	5.482	2.471	Yes
$\eta_3 - \eta_4$	0.2	0.2641	0.6981	0.2375	0.726	-1.144	1.715	2.859	No

TABLE 6.2. Using **residual** bootstrap: False coverage-statement rate (FCR) - Adjusted BH-Selected CIs for selected parameters indicated by **the percentile bootstrap CI**; All confidence intervals above show significance against $H_0 : \eta_i = \eta_{i'}$

	TRUE	mean	std	est	p-value	CI lower	CI upper	CI length	reject?
$\eta_1 - \eta_2$	-1.2	-2.0383	0.7696	-2.0430	0.001	-3.549	-0.402	3.147	Yes
$\eta_1 - \eta_3$	2.5	1.9758	0.7373	2.0172	0.003	0.457	3.393	2.936	Yes
$\eta_1 - \eta_4$	2.7	2.2372	0.7500	2.2547	$< 10^{-5}$	0.695	3.825	3.130	Yes
$\eta_2 - \eta_3$	3.7	4.0141	0.7400	4.0602	$< 10^{-5}$	2.498	5.440	2.942	Yes
$\eta_2 - \eta_4$	3.9	4.2756	0.7255	4.2977	$< 10^{-5}$	2.776	5.802	3.026	Yes
$\eta_3 - \eta_4$	0.2	0.2615	0.7381	0.2375	0.709	-1.187	1.738	2.925	No

TABLE 6.3. Using **parametric** bootstrap: False coverage-statement rate (FCR) - Adjusted BH-Selected CIs for selected parameters indicated by **surrogate null distribution**; All confidence intervals above show significance against $H_0 : \eta_i = \eta_{i'}$

	TRUE	mean	std	est	p-value	CI lower	CI upper	CI length	reject?
$\eta_1 - \eta_2$	-1.2	-2.0445	0.6610	-2.0430	0.001	-3.359	-0.652	2.707	Yes
$\eta_1 - \eta_3$	2.5	1.9618	0.7426	2.0172	0.006	0.320	3.464	3.144	Yes
$\eta_1 - \eta_4$	2.7	2.2259	0.6320	2.2547	$< 10^{-5}$	0.939	3.481	2.542	Yes
$\eta_2 - \eta_3$	3.7	4.0063	0.7363	4.0602	$< 10^{-5}$	2.481	5.403	2.922	Yes
$\eta_2 - \eta_4$	3.9	4.2704	0.6017	4.2977	$< 10^{-5}$	3.011	5.482	2.471	Yes
$\eta_3 - \eta_4$	0.2	0.2641	0.6981	0.2375	0.693	-1.144	1.715	2.859	No

TABLE 6.4. Using **residual** bootstrap: False coverage-statement rate (FCR) - Adjusted BH-Selected CIs for selected parameters indicated by **surrogate null distribution**; All confidence intervals above show significance against $H_0 : \eta_i = \eta_{i'}$

testing results are shown in Figure 6.5. The results are consistent to the true. They indicate the calculated p-value is nearly 0 when there is interaction, while the one is relatively large when there is no interaction effect. The testing results are consistent to the true setting and our proposed procedures are validated.

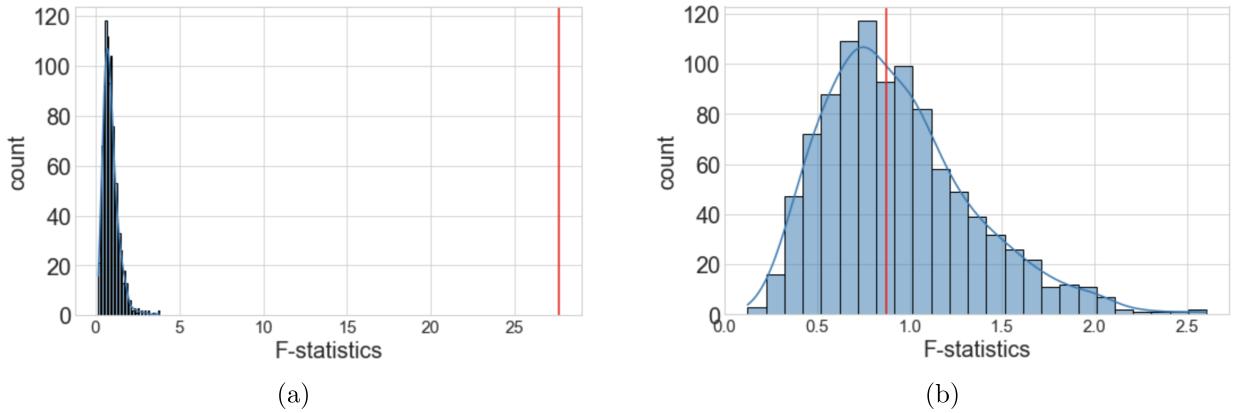
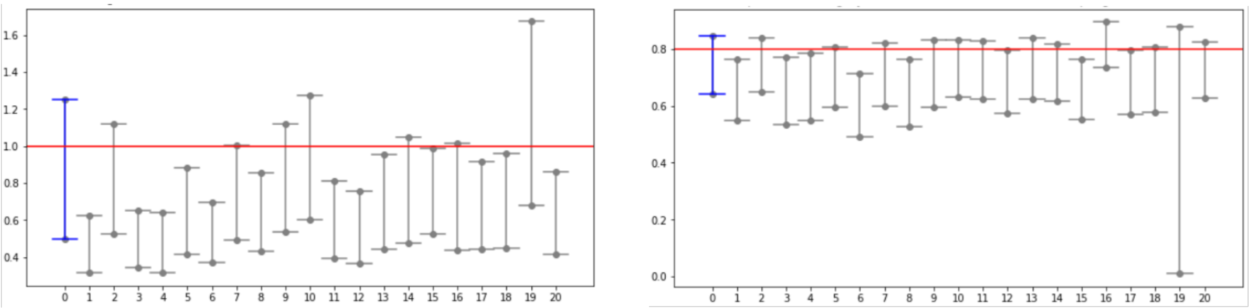


FIGURE 6.5. observed F-statistics (red line) v.s. the null bootstrap-based F-distribution: (a) Simulation 1: H_0 is not true and p-value $< 10^{-5}$; (b) Simulation 2: H_0 is true and p-value $= 0.493$.

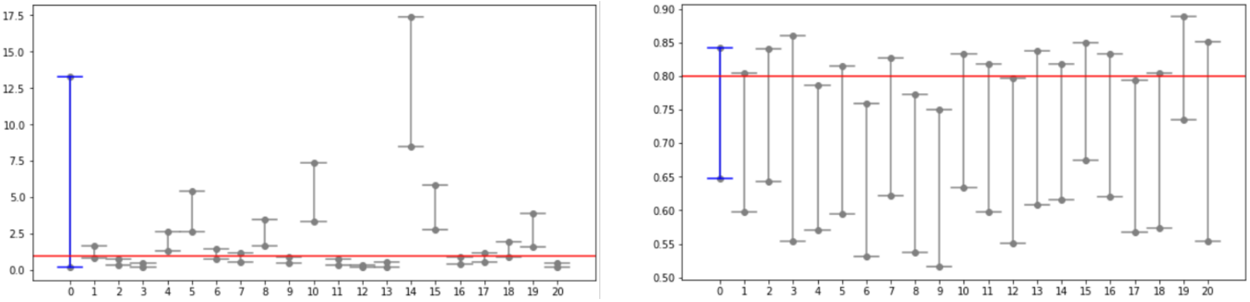
6.3.2. Dependent observations with t-distributed assumption. We consider a more complicated case when observations are correlated and follow a t-distribution. Using t-distribution for inference could be more realistic because it helps capture extreme values that might be commonly observed in practice. Figure 6.6 and Figure 6.7 indicate the necessity of applying t-bootstrap procedures for inference of variance parameters. With the number of the degree of freedom is relatively large (i.e., 200), data tends to Gaussian and thus we can see the normal assumption works well in bootstrap steps for both variance coefficients, as the coverages of 20 bootstrap intervals are comparable to the one under the true distribution. However, when data has heavy tails (i.e., the number of degree of freedom being 3), it is clear that the Gaussian-type bootstrap intervals fail to capture the variability of σ_a^2.

For simplicity, we generate data using the simplified covariance structure that involves two covariance parameters, σ_a^2, ρ. In this case, residual bootstrap becomes less useful, and therefore the aforementioned parametric bootstrap procedures are conducted for the t-distributed data in a similar way as we do for Gaussian data. The only difference is that we generate samples from an assumed t-distribution instead of a Gaussian distribution. Since the number of degree of freedom is unknown, we start with the selection of the degree of freedom. And then we compare the characteristics of the t-bootstrap distributions to the sampling distribution.



(a) (b)

Figure 6.6. When data is following t-distribution with the number of degree of freedom being 200: 95% interval of true distribution is in blue, 20 gaussian bootstrap intervals are in grey and the true value is marked as a red horizontal line.

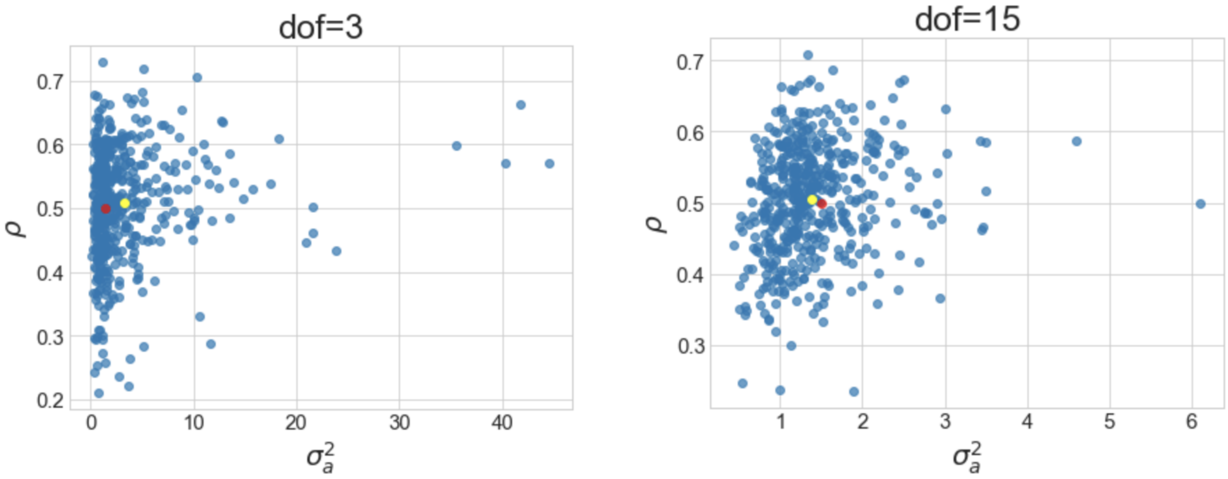


(a) (b)

Figure 6.7. When data is following t-distribution with the number of degree of freedom being 3: 95% interval of true distribution is in blue, 20 gaussian bootstrap intervals are in grey and the true value is marked as a red horizontal line.

Given a set of number of degree of freedom, $(3, 4, 15, 40, 300)$, we compute the coverage performance of each candidate. 10% of data are chosen as test set and we set $\tilde{C} = 0.8$. The selection will be conducted for 5 times to ensure of stability. As we mentioned before, the procedures may be likely to choose a smaller number of degree of freedom because the selection is dependent upon the prediction intervals by t-distributions which converge to Gaussian distributions with a large number degree of freedom. And thus the prediction intervals could be narrower and hard to cover the observations when the number of degree of freedom becomes big.

In the following simulation study, we are suggested with 3 as the optimal number of degree of freedom when the true number is 15. Then we conduct the bootstrap sampling steps given t-distribution with the number of degree of freedom as 3 and 15, respectively. Figure 6.8 indicate the bootstrap sampling distributions of two variance parameters in terms of two numbers of degree of freedom. When we use $\nu_d = 3$, we observe more extreme values in σ_a^2 than $\nu_d = 15$. This is reasonable due to the characteristics of long tails of t-distributions. In comparison, the distributions of ρ with two numbers of degree of freedom are very similar. By comparing the true values, we can see our selected ν_d performs well as the samples surround the true and the sample mean is close to the true (Figure 6.8(a)). The results demonstrate the robustness of the bootstrap sampling procedures given t-distributed data, even with an inaccurate degree of freedom.



(a)

(b)

FIGURE 6.8. Scatter plots of ρ v.s. σ_a^2 with two numbers of the degree of freedom (the red points are the true; the yellow points are the sample mean) (a) selected number of degree of freedom as 3; (b) true number of degree of freedom as 15.

6.4. Discussion

We develop an extended two-factor ANOVA framework, which preserves the effectiveness and flexibility of the simplified version developed in 5, and succeed in characterizing the treatment, time and interaction effect. The construction and analysis show the valuable capability of this advanced extended structure and techniques to answer more practical crucial experimental questions about the interaction presence and the comparative treatment effects where the data can be extremely limited. Moreover, we consider a more comprehensive variance structure that is able to account for the correlation over time within replicates. In addition, this structure is shown to be effective in explaining two different sources of variation.

Furthermore, we demonstrate this extended framework still works well in modeling and making inferences where data commonly is limited in biological experiments. The numerical experiments presented indicate the well-designed resampling procedures help provide a good description of the variability, which is very informative for experimenters to formulate and validate a robust design, according to the length of the experiments and number of replicates. We also noticed that when the data is correlated in terms of time, the choice of distribution assumptions in the parametric bootstrap process plays a key role. It is found Guassian-bootstrap fails to characterize the whole variability of the variance parameters when the data is likely to be t-distributed, and thus t-bootstrap leads to more conservative results in general. Hence, we need to be cautious about the distribution assumed in this case before making any inferences. In particular, if one is very confident about the Gaussian assumption of the data, it is recommended to use Gaussian-bootstrap for more accurate results. Otherwise, t-bootstrap is suggested as a better choice.

One potential future work is to extend the current framework by incorporating random effects. This is motivated by the fact that the shapes of trajectories given different treatments can be very different. One simple solution to this is incorporating a random intercept to allow for the baseline being random. This construction is simple but only considers the baseline variation across treatments. Another way is to consider the coefficients α_{il}'s to be random effects. With the non-decreasing constraints of the shape, we can opt to model the difference in coefficients and assume the Inverse Gamma as its prior distribution. Then coefficients can be estimated through a Bayesian analysis by running a MCMC algorithm.

CHAPTER 7

Discussion

The real data currently becomes more challenging with complex structures and practical limitations, where the traditional statistical methods are not applicable. It is essential to integrate information from complicated data systems into advanced computation for better interpretability and higher efficiency. Advanced data-driven approaches to analyzing real data are imperative to characterize and deliver the information behind the data. This realistic and essential problem inspires us to develop robust and efficient data-driven computational technologies to interpret the data system better.

This dissertation consists of work on several areas where data-driven methods are in need to answer real-world questions of interest, that traditional methods fail. We study multiple different data sets where each of them inherits different characteristics and practical limits. Based on the scientific goals and practical constraints, we develop computational approaches according to their specific characteristics and scientific goals. All work shows the merits that data-driven approaches are fundamental and reliable to taking advantage of the information from the complex system and resolving the practical questions.

The first field we work on in this dissertation is about developing an effective computational framework of data science applied to color science. Specifically, we develop a solid computational approach to analyzing color images, so that the regions of interest (with either regular or irregular bounds) can be identified according to color. One challenging problem is the large sample size of pixels in one digital image because it makes computation inefficient. By taking advantage of color consistency and unsupervised learning algorithms, the efficiency can be greatly improved for identifying targeted regions that contain fruitful information in terms of spatial and color that can be utilized to answer many practical questions. In Chapter 2, a set of small purple dots are separated from the yellow testing paper, and spatial patterns are characterized for a 2D uniform test. In addition, the identification of color regions of interest indicates the merits of investigating digital paintings of *Sunflowers in a vase* by van Gogh in Chapter 3. We detect the difference between the London version (F454) and its two repetitions (Tokyo version, F457 and Amsterdam version F458) as well as the markers of aging, both of which convey valuable messages in uncovering van Gogh's intents in color language and the aging patterns. Admittedly, the analyses involve some subjective work possibly due to the color distortions by cameras and resolution. This potential issue can be mitigated if there is a chance to work with related researchers and museum curators for image manipulation. Additionally, it could be inaccurate to differentiate target regions from the others in an image if the colors of both are too similar. In this case, we need to work closely with experts to understand the domain knowledge so that we can make use of the wealth of color information for algorithm derivations.

Besides, we build an entropy-based scheme to identify significant factors of Alzheimer's disease (AD) by investigating the heterogeneity of Alzheimer's Disease Neuroimaging Initiative (ADNI) data in Chapter 4. By applying the Categorical Exploratory Data Analysis (CEDA) scheme, we study the association by evaluating the conditional entropy-based associative patterns between the response (time-to-event) and covariates. The analysis presented indicates effective and informative results about how major factors impact the process from Mild Cognitive Impairment (MCI) to AD. One advantage is that the algorithm is efficient and simple to uncover the heterogeneity intrinsically by measuring the conditional entropy (CE) in different subsets. Notably, the whole computational process has fewer assumptions and is able to identify more major factors than Cox PH analysis. Moreover, our framework allows for any form of interaction of factors, which is, in contrast, hard to be captured by traditional methods such as the Cox PH model or linear regression.

Finally, we develop a flexible ANOVA framework that works well for limited longitudinal data in biological experiments. The main goals are modeling the mean trajectories across treatments, and making inferences with limited data. To achieve the goals, we overcome the following challenges: first, the mean curves have non-decreasing constraints because of natural growth shape; second, the large sample theory is not reliable with a significantly small sample size. In terms of those issues, we derive a solid ANOVA framework with linear constraints to model the mean trajectories and construct bootstrap-based resampling strategies for p-value computation and confidence interval construction. Within this framework, (non)linear forms of the trajectory, as the parameters of interest, can be studied and compared to select the best experimental condition. The model shows in addition the capability to quantify the treatment effect as well as the interaction effect between treatment and time. This makes it possible to answer the questions of the presence of interaction effects and how treatment effects differ across conditions. More importantly, correlated observations over time, which are commonly seen in longitudinal repeated measurement data, can be easily incorporated into our framework to quantify the sources of variation. During the resampling process, the choice of bootstrap is crucial and we should be cautious for reliable results. The parametric bootstrap with Gaussian or t-distribution is suggested when one is confident about the distribution of the observations. Otherwise, it would be more appropriate to adopt residual bootstrap for conservative results. One limitation is that the current ANOVA framework may fail to capture the different shapes of trajectories across treatments. In particular, it is possible that one trajectory preserves a distinct non-decreasing shape from others, where our current framework is not capable to characterize this variation. One of the future works is introducing random effects in this ANOVA framework. With appropriate priors assumed for the coefficients, it is promising to ensure the non-decreasing pattern and quantify the variation in the shapes of curves by applying a MCMC algorithm.

APPENDIX A

Appendix of Chapter 5

A.1. Details on resampling strategies

Nonparametric bootstrap with replicates When the number of replicates n is relatively but not extremely small, we construct the confidence interval for one parameter (e.g., the difference between a fixed pair of treatments) as follows, using a nonparametric bootstrap procedure. We resample with replacement randomly, samples of size n, from the collection of observations $\{Y_{ijk}\}_{k=1}^{n}$ for each treatment-time pair (i, j). We repeat this process B times and denote $\{Y_{ijk}^{*b}\}_{k=1}^{n}$ as the b-th sample corresponding to pair (i, j). Then we fit the model given by equations (1), (2), (4) and (5) and obtain a set of estimated parameters we are interested, denoted by $\hat{\Theta}_1^{*b}, \ldots, \hat{\Theta}_I^{*b}$.

Parametric bootstrap with replicates (without prior information on variability) The procedures for parametric bootstrap method are as follows:

- Calculate the sample variance for the i-th treatment at time t_j based on n_i replicates (for balanced data: $n_i = n$): $\hat{s}_{ij}^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (Y_{ijk} - \hat{\mu}_i(t_j))^2$
- Compute the pooled variance $\hat{s}_j^2 = \frac{\sum_i^I (n_i - 1)\hat{s}_{ij}^2}{\sum_i^I (n_i - 1)}$
- Simulate parametric bootstrap samples $Y_{ijk}^{*pb} = \hat{\mu}_i(t_j) + \epsilon_{ijk}^{*pb}$ where $\epsilon_{ijk}^{*pb} \sim N(0, \hat{s}_j^2)$.

Parametric bootstrap without replicates (with prior information on variability) Note that here we make use of pooled variance $S\hat{D}2_{.j}$ for all treatments at time j based on $\{SD_{ij}^2\}_{i,j}$. That is, $S\hat{D}2_{.j} = \frac{1}{I} \sum_i SD_{ij}^2$. We use two different strategies to generate bootstrap samples. One is using normal distribution while the other one is t-distribution, which allows for more extreme values.

- **Gaussian noise:** At each time $b = 1, \ldots, B$, we generate samples by $\mu_{ij}^{*b} = \hat{\mu}_{ij} + \epsilon_{ij}^{*b}$, where $\epsilon_{ij}^{*b} \sim N(0, S\hat{D}2_{.j})$. We use $\{\mu_{ij}^{*b}\}_{ij}$ for model fitting, leading to $\{\hat{\Theta}_i^{*b}\}_{b=1}^{B}$.
- ***t*-distributed noise:** Instead of assuming noise approximately follows normal distribution, we use t-distribution to allow more extreme values. We obtain the scale parameter

by $S\hat{D}2_{.j} = \sigma_j^2 \times \frac{df}{df-2}$ where df is the degree of freedom of t-distribution; $S\hat{D}2_{.j}$ is the measurement error (pooled variance). Then we generate noise which follows $\sigma_j t_{df}$. Here we assume $df = 3$, which means the variance is finite.

Residual bootstrap with or without replicates The general route of generating residual bootstrap samples with or without replicates (i.e., $n \geq 1$)is as follows:

(1) Suppose $\{Y_{ijk}\}$ as the raw data and obtain the $\hat{\mu}_i(t_j)$;

(2) Obtain $\hat{e}_{ijk} = Y_{ijk} - \hat{\mu}_i(t_j)$, which is biased due to the framework;

(3) For each j, obtain e^{*b}_{ijk} resampled from $\{\hat{e}_{ijk} - \bar{e}_{.j.}\}$ for $i = 1, \ldots, I$ and $k = 1, \ldots, n$ where $\bar{e}_{.j.} = \frac{1}{I\times n}\sum_i \sum_k \hat{e}_{ijk}$. Note that $\{\hat{e}_{ijk} - \bar{e}_{.j.}\}$ will sum to 0 for each j.

(4) Generate residual bootstrap samples: $Y^{*b}_{ijk} = \hat{\mu}_i(t_j) + e^{*b}_{ijk}$.

Nonparametric bootstrap based on the null distribution:

In each of the aforementioned hypothesis testing problems, the estimation problem under the null hypothesis remains a quadratic programming problem, with linear equality or inequality constraints on the parameters. In each case, the key parameters involved can be expressed in the form

$$\sum_{l=1}^{L} r_l \alpha_{il} \tag{A.1}$$

where r_l's are some constants and α_{il} is the coefficient of $\mu_i(t)$ associated with the l-th B-spline $B_l(t)$. Since the estimation problem is a quadratic programming problem, we can use the fitted trajectories to construct surrogate bootstrap data and then make use of this to compute the null distribution of the test statistics, and thereby obtain the p-value of the test. The surrogate data are computed as

$$\tilde{Y}_{ijk} = e^{b}_{ijk} + \tilde{\mu}_i(t_j) \tag{A.2}$$

where e^b_{ijk} are sampled by bootstrap procedures based on observational residuals $\{Y_{ijk} - \hat{\mu}_i(t_j)\}$, $\tilde{\mu}_i$ is the estimate of μ_i under the constraints imposed by the null hypothesis. The detailed algorithm using the null distribution is shown at the end of Section A.2.

A.2. Computation of p-value associated with tests of hypothesis

p-value computation by percentile-t bootstrap: Suppose we are performing pairwise tests for equality of the Θ_i's, where Θ_i is some functional of μ_i , by doing percentile-t bootstrap. The $100(1-\alpha)\%$ bootstrap confidence intervals for $\delta_{ij} = \Theta_i - \Theta_j$ are of the form

$$[\hat{\delta}_{ij} - q_{1-\alpha/2}\hat{\sigma}_{ij}, \hat{\delta}_{ij} - q_{\alpha/2}\hat{\sigma}_{ij}] \tag{A.3}$$

where $\delta_{ij} = \Theta_i - \Theta_j$, $\hat{\sigma}_{ij}$ is the estimated standard error of $\hat{\delta}_{ij}$ and q_p is the p-th quantile of the bootstrap distribution of the t-statistics

$$(\hat{\delta}^{*b}_{ij} - \hat{\delta}_{ij})/\hat{\sigma}^{*b}_{ij} \tag{A.4}$$

where $\hat{\delta}^{*b}_{ij}$ is the bootstrap estimate of δ_{ij} and $\hat{\sigma}^{*b}_{ij}$ is the bootstrap estimate of its standard error. Then, we may define a bootstrap p-value for testing $H^{ij}_0 : \Theta_i = \Theta_j$ to be the largest value of α such that the aforementioned $100(1-\alpha)\%$ bootstrap confidence interval contains 0. Note that $\hat{\sigma}_{ij}$ and $\hat{\sigma}^{*b}_{ij}$ can be estimated by using "double bootstrap" procedure described below (assuming only one replicate):

(1) Generate bootstrap samples $\{Y^{*b}_{ij}\}^B_{b=1}$, where B is large (say $B = 1000$) and Y^{*b}_{ij} is the b-th bootstrap measurement on i-th treatment at time t_j.

(2) Let $\hat{\Theta}^{*b}_i$ be the corresponding bootstrap estimate of Θ_i, and correspondingly, $\hat{\delta}^{*b}_{ik} = \hat{\Theta}^{*b}_i - \hat{\Theta}^{*b}_k$.

(3) Estimate σ^2_{ik} as follows:

$$\hat{\sigma}^2_{ik} = (B-1)^{-1}\sum_{b=1}^{B}(\hat{\delta}^{*b}_{ik} - \bar{\delta}^{*}_{ik})^2$$

where $\bar{\delta}^*_{ik} = B^{-1}\sum_{b=1}^{B}\hat{\delta}^{*b}_{ik}$ is a bootstrap estimate of $E(\hat{\delta}_{ik})$.

(4) In the percentile-t bootstrap procedure, we compute a bootstrap estimate of $\hat{\sigma}_{ik}$, namely, $\hat{\sigma}^{*b}_{ik}$, corresponding to the b-th bootstrap sample used for inference. This will involve repeating the procedure as described above, but now treating the data for the b-th bootstrap sample (Y^{*b}_{ij}) as the raw data.

(5) The procedure for computing $\hat{\sigma}^{*b}_{ik}$ is actually a so-called "double bootstrap" procedure. Again, these bootstrap samples for each b will have to be generated independently. So, this is clearly a computationally intensive procedure.

p-value computation by percentile bootstrap The percentile bootstrap is less computationally intensive since it does not require computation of $\hat{\sigma}^{*b}_{ik}$ for each bootstrap sample, and so there is no need for the second layer of bootstrap. We only need to find the bootstrap p-value to be the largest value of α such that the $100(1-\alpha)\%$ bootstrap confidence interval contains 0.

p-value computation by nonparametric bootstrap based on the null distribution Here are the detailed steps of the bootstrap procedure for approximating the sampling distribution under the null hypothesis of no difference across the different treatments:

(1) Use the current method to compute $\hat{\mu}_i(t)$ for the different treatments.

(2) Obtain the fitted curves $\tilde{\mu}_i(t)$ for the different treatments from the original data by imposing the additional constraint on the coefficients (α_{il}) that

$$\Theta_1 = \cdots = \Theta_I, \text{ that is }, \sum_l r_l\alpha_{1l} = \cdots = \sum_l r_l\alpha_{Il}$$

(3) Create surrogate bootstrap data $\tilde{Y}^b_{ijk} = e^b_{ijk} + \tilde{\mu}_i(t_j), i = 1,\ldots,I, j = 1,\ldots,J$, where e^b_{ijk} are generated based on parametric (t-distribution) or non-parametric (sampled from $= Y_{ijk} - \hat{\mu}_i(t_j)$) bootstrap sampling distribution. $\{\{\tilde{Y}_{ijk}\}^b\}$ are the bootstrap samples under the null hypothesis.

(4) Obtain bootstrap estimates $\{\hat{\mu}^{*b}_i(t)\}^B_{b=1}$. This leads us to a set of bootstrap estimates of Θ_i, $\{\Theta^{*b}_i\}^B_{b=1}$.

(5) Use the bootstrap sampling distribution computed in Step 4 to obtain the p-values for testing $\Theta_i = \Theta_{i'}$ for all pairs (i, i').

Notice that Step 2 of the above procedure imposes linear equality constraints on the parameters (α_{il}), along with the original monotonicity and non-negativity constraints, and therefore the resulting least squares problem can still be solved by quadratic programming. Thus, Step 2 obtains least squares estimates of $\mu_1, \cdots, \mu_I$, under the monotonicity constraint and the additional requirement (null hypothesis) that $\Theta_1 = \cdots = \Theta_I$. This is clearly less stringent than requiring that μ_i's are all equal. In step 5, for example, if $\tau_i = \max_t \mu_i(t)$, then the p-value for the test will be computed as the fraction of times $|\hat{\tau}_i^* - \hat{\tau}_k^*| > |\hat{\tau}_i - \hat{\tau}_k|$. Here, $\hat{\tau}_i = \max_t \hat{\mu}_i(t)$ (with $\hat{\mu}_i$ computed in Step 1) and $\hat{\tau}_i^{*b} = \max_t \hat{\mu}_i^{*b}(t)$ (with $\hat{\mu}_i^{*b}$ computed in Step 4).

A.3. Simultaneous inference

(1) Sort the p-values used for testing the m hypotheses regarding the parameters, $p_{(1)} \leq \cdots \leq p_{(M)}$, where M is the number of tests.

(2) Calculate $R = \max\{i : p_{(i)} \leq \frac{i}{M}\alpha\}$ where α is the significance level.

(3) Select R parameters for which $p_{(i)} \leq R\frac{\alpha}{M}$, corresponding to the rejected hypothesis.

(4) Construct a $1 - R\frac{\alpha}{m}$ CI for each parameter selected.

A.4. Figures

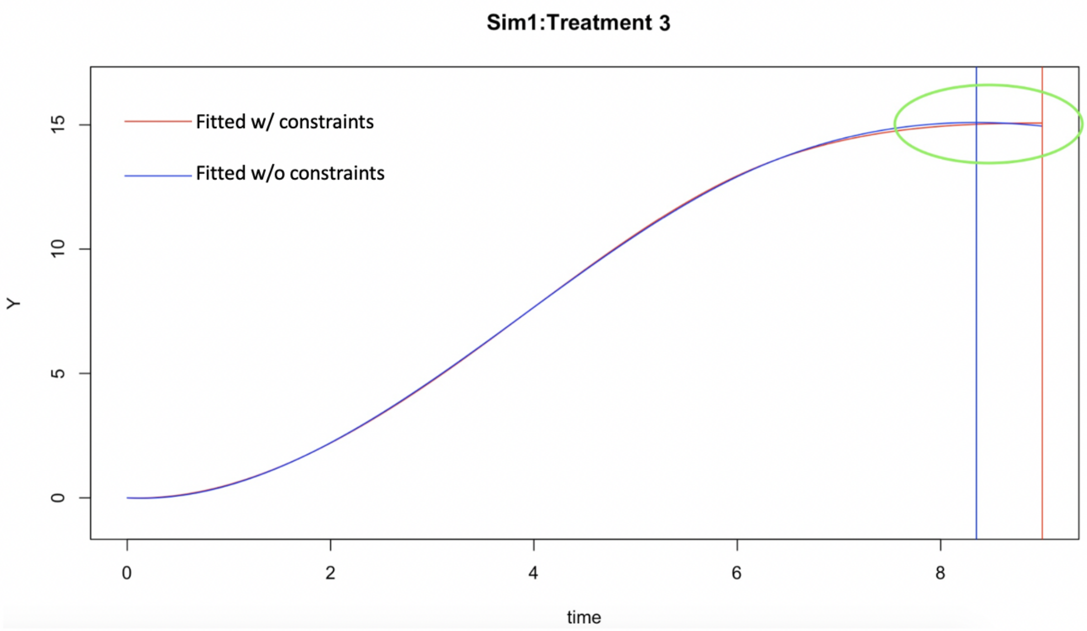


FIGURE A.1. In the simulation study of Treatment 3: blue curve represents unconstrained estimation while the red curve is constrained estimation. Decreasing pattern shown at the end of time point in treatment 3; the peak value of unconstrained estimation is shown by blue vertical line while the red vertical line for the peak of constrained one.

A.5. Tables

	p-value by null bootstrap distribution	CI lower	CI upper	CI length	mean	sd	est
$\hat{\tau}_A^{bs} - \hat{\tau}_H^{bs}$	$< 10^{-5}$	-28.082	-9.975	18.107	-19.298	3.624	-19.186
$\hat{\tau}_B^{bs} - \hat{\tau}_H^{bs}$	$< 10^{-5}$	-25.337	-8.495	16.842	-17.707	3.213	-18.115
$\hat{\tau}_C^{bs} - \hat{\tau}_H^{bs}$	$< 10^{-5}$	-26.293	-9.439	16.853	-18.197	3.379	-18.347
$\hat{\tau}_D^{bs} - \hat{\tau}_H^{bs}$	$< 10^{-5}$	-20.150	-3.602	16.547	-12.414	3.191	-12.679
$\hat{\tau}_E^{bs} - \hat{\tau}_H^{bs}$	$< 10^{-5}$	-24.672	-6.965	17.708	-15.496	3.489	-15.233
$\hat{\tau}_F^{bs} - \hat{\tau}_H^{bs}$	$< 10^{-5}$	-29.974	-11.644	18.330	-20.539	3.677	-20.223
$\hat{\tau}_G^{bs} - \hat{\tau}_H^{bs}$	$< 10^{-5}$	-24.125	-7.616	16.509	-16.482	3.187	-16.687

Table A.1. Using **residual** (nonparametric) bootstrap methods: False coverage-statement rate (FCR) - Adjusted BH-Selected CIs for selected parameters indicated by **the null bootstrap distribution**; All confidence intervals above show significance against $H_0 : \tau_i = \tau_j$

	p-value by null bootstrap distribution	CI lower	CI upper	CI length	mean	sd	est
$\hat{\tau}_A^{bs} - \hat{\tau}_D^{bs}$	0.003	-10.778	-3.105	7.674	-6.578	1.545	-6.507
$\hat{\tau}_A^{bs} - \hat{\tau}_H^{bs}$	$< 10^{-5}$	-23.481	-15.508	7.974	-19.277	1.731	-19.186
$\hat{\tau}_B^{bs} - \hat{\tau}_D^{bs}$	0.008	-9.552	-1.525	8.026	-5.301	1.525	-5.436
$\hat{\tau}_B^{bs} - \hat{\tau}_H^{bs}$	$< 10^{-5}$	-22.067	-14.185	7.882	-17.999	1.653	-18.115
$\hat{\tau}_C^{bs} - \hat{\tau}_D^{bs}$	0.006	-10.075	-1.356	8.719	-5.662	1.644	-5.668
$\hat{\tau}_C^{bs} - \hat{\tau}_H^{bs}$	$< 10^{-5}$	-22.761	-14.358	8.403	-18.361	1.709	-18.347
$\hat{\tau}_D^{bs} - \hat{\tau}_F^{bs}$	0.003	2.768	12.825	10.057	7.680	1.828	7.544
$\hat{\tau}_D^{bs} - \hat{\tau}_G^{bs}$	0.019	0.383	8.532	8.149	4.106	1.553	4.008
$\hat{\tau}_D^{bs} - \hat{\tau}_H^{bs}$	$< 10^{-5}$	-16.687	-8.478	8.209	-12.698	1.765	-12.679
$\hat{\tau}_E^{bs} - \hat{\tau}_F^{bs}$	0.022	0.121	10.278	10.157	4.859	1.877	4.990
$\hat{\tau}_E^{bs} - \hat{\tau}_H^{bs}$	$< 10^{-5}$	-20.403	-11.297	9.106	-15.519	1.925	-15.233
$\hat{\tau}_F^{bs} - \hat{\tau}_H^{bs}$	$< 10^{-5}$	-25.471	-15.836	9.635	-20.378	1.872	-20.223
$\hat{\tau}_G^{bs} - \hat{\tau}_H^{bs}$	$< 10^{-5}$	-20.893	-12.667	8.225	-16.804	1.648	-16.687

Table A.2. Using **parametric** bootstrap: False coverage-statement rate (FCR) - Adjusted BH-Selected CIs for selected parameters indicated by **the null bootstrap distribution**; All confidence intervals above show significance against $H_0 : \tau_i = \tau_j$

	p-value by	CI lower	CI upper	CI length	mean	sd	est
	percentile bootstrap CI						
$\hat{\psi}_A^{bs} - \hat{\psi}_F^{bs}$	$< 10^{-5}$	0.203	1.510	1.307	0.822	0.271	0.841
$\hat{\psi}_A^{bs} - \hat{\psi}_G^{bs}$	$< 10^{-5}$	1.163	2.082	0.919	1.698	0.204	1.714
$\hat{\psi}_A^{bs} - \hat{\psi}_H^{bs}$	$< 10^{-5}$	-1.385	-0.181	1.204	-0.796	0.260	-0.797
$\hat{\psi}_B^{bs} - \hat{\psi}_F^{bs}$	$< 10^{-5}$	0.717	1.874	1.157	1.180	0.256	1.252
$\hat{\psi}_B^{bs} - \hat{\psi}_G^{bs}$	$< 10^{-5}$	1.650	2.478	0.827	2.056	0.189	2.124
$\hat{\psi}_C^{bs} - \hat{\psi}_F^{bs}$	$< 10^{-5}$	0.571	1.775	1.204	1.029	0.259	1.086
$\hat{\psi}_C^{bs} - \hat{\psi}_G^{bs}$	$< 10^{-5}$	1.501	2.349	0.848	1.905	0.193	1.959
$\hat{\psi}_D^{bs} - \hat{\psi}_F^{bs}$	$< 10^{-5}$	0.732	1.863	1.131	1.161	0.248	1.219
$\hat{\psi}_D^{bs} - \hat{\psi}_G^{bs}$	$< 10^{-5}$	1.653	2.438	0.786	2.036	0.177	2.092
$\hat{\psi}_E^{bs} - \hat{\psi}_F^{bs}$	$< 10^{-5}$	0.586	1.852	1.266	1.162	0.262	1.096
$\hat{\psi}_E^{bs} - \hat{\psi}_G^{bs}$	$< 10^{-5}$	1.496	2.409	0.914	2.038	0.203	1.968
$\hat{\psi}_F^{bs} - \hat{\psi}_G^{bs}$	$< 10^{-5}$	0.267	1.209	0.942	0.876	0.210	0.872
$\hat{\psi}_F^{bs} - \hat{\psi}_H^{bs}$	$< 10^{-5}$	-2.274	-1.019	1.254	-1.619	0.267	-1.638
$\hat{\psi}_G^{bs} - \hat{\psi}_H^{bs}$	$< 10^{-5}$	-2.875	-1.977	0.898	-2.495	0.201	-2.511

Table A.3. Using **residual** bootstrap method: False coverage-statement rate (FCR) - Adjusted BH-Selected CIs for selected parameters indicated by **the percentile bootstrap CIs**; All confidence intervals above show significance against $H_0 : \psi_i = \psi_j$

	p-value by null bootstrap distribution	CI lower	CI upper	CI length	mean	sd	est
$\hat{\psi}_A^{bs} - \hat{\psi}_B^{bs}$	0.004	-0.644	-0.174	0.471	-0.401	0.110	-0.410
$\hat{\psi}_A^{bs} - \hat{\psi}_D^{bs}$	0.008	-0.594	-0.158	0.436	-0.369	0.106	-0.378
$\hat{\psi}_A^{bs} - \hat{\psi}_E^{bs}$	0.030	-0.611	-0.034	0.578	-0.282	0.137	-0.255
$\hat{\psi}_A^{bs} - \hat{\psi}_F^{bs}$	$< 10^{-5}$	0.507	1.127	0.620	0.823	0.135	0.841
$\hat{\psi}_A^{bs} - \hat{\psi}_G^{bs}$	$< 10^{-5}$	1.501	1.894	0.393	1.698	0.095	1.714
$\hat{\psi}_A^{bs} - \hat{\psi}_H^{bs}$	0.002	-1.055	-0.568	0.486	-0.809	0.123	-0.797
$\hat{\psi}_B^{bs} - \hat{\psi}_F^{bs}$	$< 10^{-5}$	0.940	1.488	0.549	1.224	0.123	1.252
$\hat{\psi}_B^{bs} - \hat{\psi}_G^{bs}$	$< 10^{-5}$	1.914	2.261	0.347	2.099	0.081	2.124
$\hat{\psi}_B^{bs} - \hat{\psi}_H^{bs}$	0.006	-0.628	-0.205	0.423	-0.407	0.112	-0.387
$\hat{\psi}_C^{bs} - \hat{\psi}_F^{bs}$	$< 10^{-5}$	0.763	1.312	0.549	1.055	0.123	1.086
$\hat{\psi}_C^{bs} - \hat{\psi}_G^{bs}$	$< 10^{-5}$	1.717	2.096	0.379	1.929	0.084	1.959
$\hat{\psi}_C^{bs} - \hat{\psi}_H^{bs}$	0.002	-0.813	-0.374	0.439	-0.577	0.112	-0.552
$\hat{\psi}_D^{bs} - \hat{\psi}_F^{bs}$	$< 10^{-5}$	0.885	1.448	0.563	1.192	0.122	1.219
$\hat{\psi}_D^{bs} - \hat{\psi}_G^{bs}$	$< 10^{-5}$	1.885	2.212	0.327	2.067	0.078	2.092
$\hat{\psi}_D^{bs} - \hat{\psi}_H^{bs}$	0.002	-0.672	-0.246	0.426	-0.439	0.112	-0.419
$\hat{\psi}_E^{bs} - \hat{\psi}_F^{bs}$	$< 10^{-5}$	0.800	1.426	0.626	1.105	0.143	1.096
$\hat{\psi}_E^{bs} - \hat{\psi}_G^{bs}$	$< 10^{-5}$	1.784	2.241	0.458	1.979	0.110	1.968
$\hat{\psi}_E^{bs} - \hat{\psi}_H^{bs}$	0.008	-0.778	-0.236	0.542	-0.527	0.138	-0.542
$\hat{\psi}_F^{bs} - \hat{\psi}_G^{bs}$	$< 10^{-5}$	0.629	1.145	0.516	0.875	0.112	0.872
$\hat{\psi}_F^{bs} - \hat{\psi}_H^{bs}$	$< 10^{-5}$	-1.906	-1.342	0.564	-1.631	0.135	-1.638
$\hat{\psi}_G^{bs} - \hat{\psi}_H^{bs}$	$< 10^{-5}$	-2.696	-2.323	0.373	-2.506	0.097	-2.511

Table A.4. Using **parametric** bootstrap method: False coverage-statement rate (FCR) - Adjusted BH-Selected CIs for selected parameters indicated by **the percentile bootstrap CIs**; All confidence intervals above show significance against $H_0 : \psi_i = \psi_j$

	CI lower	CI upper	CI length	mean	sd	est
$\hat{\theta}^{bs}_A$	3.493	3.849	0.355	3.642	0.096	3.619
$\hat{\theta}^{bs}_B$	2.758	2.978	0.220	2.841	0.064	2.813
$\hat{\theta}^{bs}_C$	2.958	3.238	0.280	3.081	0.077	3.063
$\hat{\theta}^{bs}_D$	2.513	2.678	0.165	2.587	0.046	2.588
$\hat{\theta}^{bs}_E$	3.138	3.463	0.325	3.278	0.081	3.268
$\hat{\theta}^{bs}_F$	4.174	4.484	0.310	4.318	0.077	4.314
$\hat{\theta}^{bs}_G$	2.673	2.878	0.205	2.763	0.055	2.748
$\hat{\theta}^{bs}_H$	2.578	2.743	0.165	2.659	0.042	2.658
$\hat{\theta}^{bs}_A - \hat{\theta}^{bs}_B$	0.591	1.021	0.430	0.801	0.116	0.806
$\hat{\theta}^{bs}_A - \hat{\theta}^{bs}_C$	0.315	0.801	0.485	0.561	0.122	0.556
$\hat{\theta}^{bs}_A - \hat{\theta}^{bs}_D$	0.876	1.271	0.395	1.055	0.107	1.031
$\hat{\theta}^{bs}_A - \hat{\theta}^{bs}_E$	0.130	0.631	0.501	0.364	0.125	0.350
$\hat{\theta}^{bs}_A - \hat{\theta}^{bs}_F$	-0.911	-0.440	0.470	-0.676	0.122	-0.696
$\hat{\theta}^{bs}_A - \hat{\theta}^{bs}_G$	0.686	1.086	0.400	0.879	0.110	0.871
$\hat{\theta}^{bs}_A - \hat{\theta}^{bs}_H$	0.806	1.216	0.410	0.983	0.105	0.961
$\hat{\theta}^{bs}_B - \hat{\theta}^{bs}_C$	-0.430	-0.060	0.370	-0.240	0.102	-0.250
$\hat{\theta}^{bs}_B - \hat{\theta}^{bs}_D$	0.130	0.415	0.285	0.254	0.079	0.225
$\hat{\theta}^{bs}_B - \hat{\theta}^{bs}_E$	-0.631	-0.245	0.385	-0.437	0.103	-0.455
$\hat{\theta}^{bs}_B - \hat{\theta}^{bs}_F$	-1.662	-1.281	0.380	-1.477	0.102	-1.502
$\hat{\theta}^{bs}_B - \hat{\theta}^{bs}_G$	-0.075	0.240	0.315	0.078	0.087	0.065
$\hat{\theta}^{bs}_B - \hat{\theta}^{bs}_H$	0.060	0.335	0.275	0.182	0.078	0.155
$\hat{\theta}^{bs}_C - \hat{\theta}^{bs}_D$	0.335	0.691	0.355	0.494	0.089	0.475
$\hat{\theta}^{bs}_C - \hat{\theta}^{bs}_E$	-0.430	0.025	0.455	-0.197	0.113	-0.205
$\hat{\theta}^{bs}_C - \hat{\theta}^{bs}_F$	-1.441	-1.006	0.435	-1.237	0.109	-1.251
$\hat{\theta}^{bs}_C - \hat{\theta}^{bs}_G$	0.145	0.506	0.360	0.318	0.094	0.315
$\hat{\theta}^{bs}_C - \hat{\theta}^{bs}_H$	0.260	0.621	0.360	0.422	0.089	0.405
$\hat{\theta}^{bs}_D - \hat{\theta}^{bs}_E$	-0.891	-0.521	0.370	-0.691	0.093	-0.681
$\hat{\theta}^{bs}_D - \hat{\theta}^{bs}_F$	-1.907	-1.552	0.355	-1.731	0.089	-1.727
$\hat{\theta}^{bs}_D - \hat{\theta}^{bs}_G$	-0.315	-0.050	0.265	-0.176	0.072	-0.160
$\hat{\theta}^{bs}_D - \hat{\theta}^{bs}_H$	-0.190	0.045	0.235	-0.072	0.063	-0.070
$\hat{\theta}^{bs}_E - \hat{\theta}^{bs}_F$	-1.246	-0.821	0.425	-1.040	0.111	-1.046
$\hat{\theta}^{bs}_E - \hat{\theta}^{bs}_G$	0.330	0.721	0.390	0.515	0.101	0.521
$\hat{\theta}^{bs}_E - \hat{\theta}^{bs}_H$	0.455	0.816	0.360	0.619	0.092	0.611
$\hat{\theta}^{bs}_F - \hat{\theta}^{bs}_G$	1.376	1.732	0.355	1.555	0.094	1.567
$\hat{\theta}^{bs}_F - \hat{\theta}^{bs}_H$	1.502	1.832	0.330	1.659	0.085	1.657
$\hat{\theta}^{bs}_G - \hat{\theta}^{bs}_H$	-0.020	0.245	0.265	0.104	0.068	0.090

TABLE A.5. Parametric Bootstrap CIs for $c = 40$; Significance (at level $\alpha = 0.05$) against $H_0 : \theta_i = \theta_j$ is in red; differences involving treatment G or H (which we are most interested in) are highlighted in blue; assuming t-distributed noise.

	CI lower	CI upper	CI length	mean	sd	est
$\hat{\gamma}_A^{bs}$	4.194	5.000	0.806	4.491	0.183	4.424
$\hat{\gamma}_B^{bs}$	3.569	3.834	0.265	3.634	0.110	3.594
$\hat{\gamma}_C^{bs}$	3.634	4.274	0.641	3.882	0.185	3.859
$\hat{\gamma}_D^{bs}$	3.589	3.939	0.350	3.722	0.117	3.719
$\hat{\gamma}_E^{bs}$	4.089	5.000	0.911	4.761	0.325	5.000
$\hat{\gamma}_F^{bs}$	4.750	5.000	0.250	4.954	0.074	5.000
$\hat{\gamma}_G^{bs}$	3.539	3.899	0.360	3.647	0.162	3.599
$\hat{\gamma}_H^{bs}$	4.374	4.965	0.591	4.514	0.123	4.459
$\hat{\gamma}_A^{bs} - \hat{\gamma}_B^{bs}$	0.455	1.401	0.946	0.857	0.212	0.831
$\hat{\gamma}_A^{bs} - \hat{\gamma}_C^{bs}$	0.090	1.196	1.106	0.609	0.255	0.566
$\hat{\gamma}_A^{bs} - \hat{\gamma}_D^{bs}$	0.370	1.311	0.941	0.769	0.217	0.706
$\hat{\gamma}_A^{bs} - \hat{\gamma}_E^{bs}$	-0.771	0.551	1.321	-0.270	0.373	-0.576
$\hat{\gamma}_A^{bs} - \hat{\gamma}_F^{bs}$	-0.796	0.035	0.831	-0.463	0.198	-0.576
$\hat{\gamma}_A^{bs} - \hat{\gamma}_G^{bs}$	0.480	1.401	0.921	0.844	0.241	0.826
$\hat{\gamma}_A^{bs} - \hat{\gamma}_H^{bs}$	-0.501	0.526	1.026	-0.022	0.223	-0.035
$\hat{\gamma}_B^{bs} - \hat{\gamma}_C^{bs}$	-0.621	0.065	0.686	-0.248	0.217	-0.265
$\hat{\gamma}_B^{bs} - \hat{\gamma}_D^{bs}$	-0.335	0.150	0.485	-0.088	0.159	-0.125
$\hat{\gamma}_B^{bs} - \hat{\gamma}_E^{bs}$	-1.426	-0.395	1.031	-1.127	0.340	-1.406
$\hat{\gamma}_B^{bs} - \hat{\gamma}_F^{bs}$	-1.426	-1.076	0.350	-1.320	0.132	-1.406
$\hat{\gamma}_B^{bs} - \hat{\gamma}_G^{bs}$	-0.265	0.220	0.485	-0.013	0.196	-0.005
$\hat{\gamma}_B^{bs} - \hat{\gamma}_H^{bs}$	-1.336	-0.616	0.721	-0.880	0.169	-0.866
$\hat{\gamma}_C^{bs} - \hat{\gamma}_D^{bs}$	-0.165	0.541	0.706	0.160	0.217	0.140
$\hat{\gamma}_C^{bs} - \hat{\gamma}_E^{bs}$	-1.316	0.000	1.316	-0.879	0.374	-1.141
$\hat{\gamma}_C^{bs} - \hat{\gamma}_F^{bs}$	-1.346	-0.666	0.681	-1.072	0.198	-1.141
$\hat{\gamma}_C^{bs} - \hat{\gamma}_G^{bs}$	-0.145	0.666	0.811	0.235	0.247	0.260
$\hat{\gamma}_C^{bs} - \hat{\gamma}_H^{bs}$	-1.076	-0.225	0.851	-0.632	0.224	-0.601
$\hat{\gamma}_D^{bs} - \hat{\gamma}_E^{bs}$	-1.406	-0.305	1.101	-1.039	0.351	-1.281
$\hat{\gamma}_D^{bs} - \hat{\gamma}_F^{bs}$	-1.406	-0.951	0.455	-1.232	0.139	-1.281
$\hat{\gamma}_D^{bs} - \hat{\gamma}_G^{bs}$	-0.200	0.345	0.546	0.075	0.202	0.120
$\hat{\gamma}_D^{bs} - \hat{\gamma}_H^{bs}$	-1.206	-0.506	0.701	-0.792	0.170	-0.741
$\hat{\gamma}_E^{bs} - \hat{\gamma}_F^{bs}$	-0.861	0.225	1.086	-0.193	0.333	0.000
$\hat{\gamma}_E^{bs} - \hat{\gamma}_G^{bs}$	0.385	1.456	1.071	1.114	0.362	1.401
$\hat{\gamma}_E^{bs} - \hat{\gamma}_H^{bs}$	-0.495	0.596	1.091	0.248	0.348	0.541
$\hat{\gamma}_F^{bs} - \hat{\gamma}_G^{bs}$	1.006	1.456	0.450	1.307	0.176	1.401
$\hat{\gamma}_F^{bs} - \hat{\gamma}_H^{bs}$	0.000	0.606	0.606	0.440	0.141	0.541
$\hat{\gamma}_G^{bs} - \hat{\gamma}_H^{bs}$	-1.326	-0.551	0.776	-0.866	0.207	-0.861

TABLE A.6. Parametric Bootstrap CIs for $c = 0.1$; Significance (at level $\alpha = 0.05$) against $H_0 : \gamma_i = \gamma_j$ is in red; differences involving treatment G or H (which we are most interested in) are highlighted in blue; assuming t-distributed noise.

Bibliography

[1] *Technical Report: Colorimetry (3rd ed.)*., CIE 15, 2004.

[2] S. S. Al-amri, N. V. Kalyankar, and D. KhamitkarS., *Image segmentation by using threshold techniques*, ArXiv, abs/1005.4020 (2010).

[3] S. Alkanaimsh, J. M. Corbin, M. J. Kailemia, K. Karuppanan, R. L. Rodriguez, C. B. Lebrilla, K. A. McDonald, and S. Nandi, *Purification and site-specific N-glycosylation analysis of human recombinant butyrylcholinesterase from Nicotiana benthamiana*, Biochemical Engineering Journal, 142 (2019), pp. 58–67.

[4] A. Altmann, L. Tian, V. W. Henderson, M. D. Greicius, and Alzheimer's Disease Neuroimaging Initiative Investigators, *Sex modifies the apoe-related risk of developing alzheimer disease*, Annals of Neurology, 75 (2014), pp. 563–573.

[5] B. A. Ardekani, N. O. Izadi, S. A. Hadid, A. M. Meftah, and A. H. Bachman, *Effects of sex, age, and apolipoprotein e genotype on hippocampal parenchymal fraction in cognitively normal older adults*, Psychiatry Research: Neuroimaging, 301 (2020), p. 111107.

[6] W. Arnold and L. Loftus, *Xanthopsia and van Gogh's yellow palette*, Eye (Lond.), 5 (1991), pp. 503–510.

[7] N. Bakker, L. Jansen, and H. Luijten, *Vincent van Gogh: A life in letters*, Thames and Hudson, London, UK, 2020.

[8] N. Bakker and C. Riopelle, *The Sunflowers in Perspective*, Amsterdam University Press, 2019, pp. 21–47.

[9] M. Barberio, E. Skantzakis, S. Sorieul, and P. Antici, *Pigment darkening as case study of in-air plasma-induced luminescence*, Science Advances, 5 (2019).

[10] Y. Benjamini and D. Yekutieli, *False discovery rate–adjusted multiple confidence intervals for selected parameters*, Journal of the American Statistical Association, 100 (2005), pp. 71–81.

[11] D. Blumer, *The illness of Vincent van Gogh*, The American journal of psychiatry, 159 (2002), pp. 519–526.

[12] S. Centeno, C. Hale, F. Carò, A. Cesarotto, N. Shibayama, J. Delaney, K. Dooley, G. van der Snickt, K. Janssens, and S. Stein, *Van Gogh's irises and roses: the contribution of chemical analyses and imaging to the assessment of color changes in the red lake pigments*, Herit. Sci., 5 (2017), p. 18.

[13] M. Chapman, S. Tang, D. Draeger, S. Nambeesan, H. Shaffer, J. Barb, S. Knapp, and J. Burke, *Genetic analysis of floral symmetry in Van Gogh's sunflowers reveals independent recruitment of CYCLOIDEA genes in the Asteraceae*, PLoS Genetics, 8 (2012), p. e1002628.

[14] T.-L. Chen, E. P. Chou, and H. Fushing, *Categorical nature of major factor selection via information theoretic measurements*, Entropy, 23 (2021).

[15] T.-L. Chen, H. Fushing, and E. P. Chou, *Multiscale major factor selections for complex system data with structural dependency and heterogeneity*, 2022.

[16] L. Chittka and J. Walker, *Do bee like Van Gogh's* Sunflowers*?*, Optics and Laser Technology, 38 (2006), pp. 323–328.

[17] E. P. Chou, T.-L. Chen, and H. Fushing, *Unraveling hidden major factors by breaking heterogeneity into homogeneous parts within many-system problems*, Entropy, 24 (2022).

[18] M. Coeckelbergh, *Can machines create art?*, Philosophy and Teachnology, 30 (2017), pp. 285–303.

[19] J. M. Corbin, B. I. Hashimoto, K. Karuppanan, Z. R. Kyser, L. Wu, B. A. Roberts, A. R. Noe, R. L. Rodriguez, K. A. McDonald, and S. Nandi, *Semicontinuous Bioreactor Production of Recombinant Butyrylcholinesterase in Transgenic Rice Cell Suspension Cultures*, Frontiers in Plant Science, 7 (2016), p. 412.

[20] D. R. Cox, *Regression models and life-tables*, Journal of the Royal Statistical Society: Series B (Methodological), 34 (1972), pp. 187–202.

[21] M. Crossley, F. Craik, and T. Salthouse, *(eds.) the handbook of aging and cognition (2nd ed.). mahwah, nj: Lawrence erlbaum, 2000.*, Canadian Journal on Aging / La Revue canadienne du vieillissement, 20 (2001), pp. 590–594.

[22] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Application*, Cambridge University Press, 1997.

[23] C. De Boor, *A Practical Guide to Splines*, Springer, 1978.

[24] P. J. Diggle, *An approach to the analysis of repeated measurements*, Biometrics, 44 (1988), pp. 959–971.

[25] P. Duarte-Guterman, A. Y. Albert, C. K. Barha, L. A. Galea, and on behalf of the Alzheimer's Disease Neuroimaging Initiative, *Sex influences the effects of apoe genotype and alzheimer's diagnosis on neuropathology and memory*, Psychoneuroendocrinology, 129 (2021), p. 105248.

[26] B. Efron, *The Two-Sample Problem with Censored Data*, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 4 (1967), pp. 831–852.

[27] B. Efron, *Better bootstrap confidence intervals*, Journal of the American Statistical Association, 82 (1987), pp. 171–185.

[28] ———, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction*, Cambridge University Press, 2010.

[29] M. V. Ellis, *Repeated measures designs*, The Counseling Psychologist, 27 (1999), pp. 552–578.

[30] G. L. Ellman, K. D. Courtney, V. Andres, and R. M. Featherstone, *A new and rapid colorimetric determination of acetylcholinesterase activity*, Biochemical Pharmacology, 7 (1961), pp. 88–95.

[31] B. Faiçal, F. Costa, G. Pessin, J. Ueyama, H. Freitas, A. Colombo, P. Fini, L. Villas, F. Osório, P. Vargas, and T. Braun, *The use of unmanned aerial vehicles and wireless sensor networks for spraying pesticides*, Journal of Systems Architecture, 60 (2014), pp. 393–404.

[32] J. Fieberg, P. Knutås, K. Hostettler, and G. Smith, *"paintings fade like flowers": Pigment analysis and digital reconstruction of a faded pink lake pigment in Vincent van Gogh's undergrowth with two figures*, Applied Spectroscopy, 71 (2017), pp. 794–808.

[33] I. Fiedler, E. Hendriks, T. Meedendorp, M. Menu, and J. Salvant, *Materials, Intention, and Evolution*, Yale University Press, New Haven, CT, USA, 2016.

[34] G. M. Fitzmaurice and C. Ravichandran, *A primer in longitudinal data analysis*, Circulation, 118 (2008), pp. 2005–2010.

[35] H. Fushing, *Semiparametric efficient inferences for lifetime regression model with time-dependent covariates*, Annals of the Institute of Statistical Mathematics, 64 (2012), pp. 1–25.

[36] H. Fushing, E. Chou, and T.-L. Chen, *Multiscale major factor selections for complex system data with structural dependency and heterogeneity*, 2022.

[37] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, Society for Industrial and Applied Mathematics, 2007.

[38] T. P. Garcia and K. Marder, *Statistical approaches to longitudinal data analysis in neurodegenerative diseases: Huntington's disease as a model.*, Current neurology and neuroscience reports, 17 (2017), p. 14.

[39] M. Geldof, A. Proaño Gaibor, F. Ligterink, E. Hendriks, and E. Kirchner, *Reconstructing van Gogh's palette to determine the optical characteristics of his paints*, Heritage Science, 6 (2018), p. 17.

[40] E. L. Glisky, *Changes in Cognitive Function in Human Aging. In D. R. Riddle (Ed.), Brain Aging: Models, Methods, and Mechanisms.*, CRC Press/Taylor & Francis, 2007.

[41] A. Gruener, *Vincent van Gogh's yellow vision*, British Journal of General Practice, 63 (2013), pp. 370–371.

[42] J. Gurevitch and S. T. Chester, *Analysis of repeated measures experiments*, Ecology, 67 (1986), pp. 251–255.

[43] P. Hanrahan and W. Krueger, *Reflection from layered surfaces due to subsurface scattering*, in Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '93, New York, NY, USA, 1993, Association for Computing Machinery, p. 165–174.

[44] E. Hendriks, A. Brokerhof, and K. van den Meiracker, *Valuing van Gogh's colours: From the past to the future*, in ICOM Committee for Conservation 18th Triennial Meeting, 2017.

[45] E. Hendriks, M. Geldof, L. Monico, D. Johnson, C. Miliani, A. Romani, C. Grazia, D. Buti, B. Brunetti, K. Janssens, G. Van der Snickt, and F. Vanmeert, *Methods and Materials of the Amsterdam Sunflowers*, Amsterdam University Press, Amsterdam, Netherlands, 2019, pp. 85–123.

[46] E. Hendriks and M. Vellekoop, *Van Gogh's Sunflowers Illuminated: Art Meets Science*, Amsterdam University Press, Amsterdam, Netherlands, 2019.

[47] C. Higgitt, G. Macaro, and M. Spring, *Methods, Materials and Condition of the London Sunflowers*, Amsterdam University Press, Amsterdam, Netherlands, 2019, pp. 49–83.

[48] B. C. Ho, N. C. Andreasen, S. Ziebell, R. Pierson, and V. Magnotta, *Long-term antipsychotic treatment and brain volumes: a longitudinal study of first-episode schizophrenia.*, Archives of general psychiatry, 68 (2011), pp. 128–137.

[49] J. Huang, T. D. Sutliff, L. Wu, S. Nandi, K. Benge, M. Terashima, A. H. Ralston, W. Drohan, N. Huang, and R. L. Rodriguez, *Expression and Purification of Functional Human α-1-Antitrypsin from Cultured Plant Cells.*, Biotechnology Progress, 17 (2001), pp. 126–133.

[50] N. Huang, J. Chandler, B. R. Thomas, N. Koizumi, and R. L. Rodriguez, *Metabolic regulation of α-amylase gene expression in transgenic cell cultures of rice (Oryza sativa L.)*, Plant Molecular Biology, 23 (1993), pp. 737–747.

[51] R. Hunt, *The Reproduction of Colour*, The Wiley-IS&T Series in Imaging Science and Technology, Wiley, 2005.

[52] M. Iwanicka, M. Sylwestrzak, A. Szkulmowska, and P. Targowski, *Methods and Techniques*, Amsterdam University Press, Amsterdam, Netherlands, 2019, pp. 207–227.

[53] S. Jain and V. Laxmi, *Color image segmentation techniques: A survey*, in Proceedings of the International Conference on Microelectronics, Computing & Communication Systems, V. Nath, ed., Singapore, 2018, Springer Singapore, pp. 189–197.

[54] R. I. Jennrich and M. D. Schluchter, *Unbalanced repeated-measures models with structured covariance matrices*, Biometrics, 42 (1986), pp. 805–820.

[55] H. W. Jensen, S. R. Marschner, M. Levoy, and P. Hanrahan, *A practical model for subsurface light transport*, in Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '01, New York, NY, USA, 2001, Association for Computing Machinery, p. 511–518.

[56] G. H. Joblove and D. Greenberg, *Color spaces for computer graphics*, SIGGRAPH Comput. Graph., 12 (1978), p. 20–25.

[57] E. L. Kaplan and P. Meier, *Nonparametric estimation from incomplete observations*, Journal of the American Statistical Association, 53 (1958), pp. 457–481.

[58] C. Kelly and J. Rice, *Monotone smoothing with application to dose-response curves and the assessment of synergism*, Biometrics, 46 (1990), p. 1071—1085.

[59] H. Keselman, J. Algina, R. K. Kowalchuk, and R. D. Wolfinger, *A comparison of recent approaches to the analysis of repeated measurements*, British Journal of Mathematical and Statistical Psychology, 52 (1999), pp. 63–78.

[60] H. J. Keselman, J. Algina, and R. K. Kowalchuk, *The analysis of repeated measures designs: A review*, British Journal of Mathematical and Statistical Psychology, 54 (2001), pp. 1–20.

[61] B. Khajehpiri, H. A. Moghaddam, M. Forouzanfar, R. Lashgari, J. Ramos-Cejudo, R. S. Osorio, B. A. Ardekani, and for the Alzheimer's Disease Neuroimaging Initiative, *Survival analysis in cognitively normal subjects and in patients with mild cognitive impairment using a proportional hazards model with extreme gradient boosting regression*, Journal of Alzheimer's Disease, 85 (2022), pp. 837–850.

[62] E. Kirchner, M. Geldof, E. Hendriks, A. Proaño Gaibor, K. Janssens, J. Delaney, I. van der Lans, F. Ligterink, L. Megens, T. Meedendorp, and K. Pilz, *Recreating van Gogh's original colors on museum displays*, in Color Imaging XXIV: Displaying, Processing, Hardcopy, and Applications, Burlingame, CA, USA, 13-17 January 2019, R. Eschbach, G. Marcu, and A. Rizzi, eds., Oxford, United Kingdom, 2019, Ingenta.

[63] E. Kirchner, I. van der Lans, F. Martínez-Verdú, and E. Perales, *Improving color reproduction accuracy of a mobile liquid crystal display*, J. Opt. Soc. Am. A, 34 (2017), pp. 101–110.

[64] C. Korenberg, *The photo-ageing behaviour of selected watercolour paints under anoxic conditions*, British Museum Technical Research Bulletin, 2 (2008), pp. 49–57.

[65] S. D. Kristjansson, J. C. Kircher, and A. K. Webb, *Multilevel models for repeated measures research designs in psychophysiology: an introduction to growth curve modeling.*, Psychophysiology, 44 (2007), pp. 728–736.

[66] J. K. Kueper, M. Speechley, and M. Montero-Odasso, *The alzheimer's disease assessment scale–cognitive subscale (adas-cog): Modifications and responsiveness in pre-dementia populations. a narrative review*, Journal of Alzheimer's Disease, 63 (2018), pp. 423–444.

[67] N. Kulkarni, *Color thresholding method for image segmentation of natural images*, International Journal of Image, Graphics and Signal Processing, 4 (2012).

[68] F. Leitenstorfer and G. Tutz, *Generalized monotonic regression based on B-splines with an application to air pollution data*, Biostatistics, 8 (2006), pp. 654–673.

[69] K. Li, R. O'Brien, M. Lutz, S. Luo, and The Alzheimer's Disease Neuroimaging Initiative, *A prognostic model of alzheimer's disease relying on multiple longitudinal measures and time-to-event data*, Alzheimer's & Dementia, 14 (2018), pp. 644–651.

[70] M. Li and P. M. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, Texts in Computer Science, Springer New York, 2009.

[71] S. Liao, K. Macharoen, K. McDonald, S. Nandi, and D. Paul, *Rice-recombinant butyrylcholinesterase (rrbche) production dataset*, Dryad, Dataset, (2021).

[72] S. Liao, K. Macharoen, K. A. McDonald, S. Nandi, and D. Paul, *Analysis of variability of functionals of recombinant protein production trajectories based on limited data*, International Journal of Molecular Sciences, 23 (2022).

[73] J. J. Locascio and A. Atri, *An overview of longitudinal data analysis methods for neurological research.*, Dementia and Geriatric Cognitive Disorders Extra, 1 (2011), pp. 330–357.

[74] O. Lockridge, *Review of human butyrylcholinesterase structure, function, genetic variants, history of use in the clinic, and potential therapeutic uses*, Pharmacology and Therapeutics, 148 (2015), pp. 34–46.

[75] T. A. Louis, *General methods for analysing repeated measures*, Statistics in Medicine, 7 (1988), pp. 29–45.

[76] K. Macharoen, K. A. McDonald, and S. Nandi, *Simplified bioreactor processes for recombinant butyrylcholinesterase production in transgenic rice cell suspension cultures*, Biochemical Engineering Journal, (2020). In press,107751.

[77] R. G. Miller, *Survival Analysis*, Wiley, 1981.

[78] U. R. Mogili and B. Deepak, *Review on application of drone systems in precision agriculture*, Procedia Computer Science, 133 (2018), pp. 502–509.

[79] R. C. Mohs, D. Knopman, R. C. Petersen, S. H. Ferris, C. Ernesto, M. Grundman, M. Sano, L. Bieliauskas, D. Geldmacher, C. Clark, and L. J. Thal, *Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's Disease Assessment Scale that broaden its scope. The Alzheimer's Disease Cooperative Study.*, Alzheimer disease and associated disorders, 11(Suppl. 2) (1997), pp. S13–21.

[80] L. Monico, E. Hendriks, M. Geldof, C. Miliani, K. Janssens, B. Brunetti, M. Cotte, F. Vanmeert, A. Chieli, G. Van der Snickt, A. Romani, and M. Melo, *Chemical Alteration and Colour Changes in the Amsterdam Sunflowers: A Focus on Geranium Lakes and Chrome Yellows*, Amsterdam University Press, Amsterdam, Netherlands, 2019, pp. 125–157.

[81] L. Monico, K. Janssens, E. Hendriks, F. Vanmeert, G. Van der Snickt, M. Cotte, G. Falkenberg, B. Brunetti, and C. Millani, *Evidence for degradation of the chrome yellows in Van Gogh's* Sunflowers*: A study using noninvasive in situ methods and synchrotron-radiation-based X-ray techniques*, Angew. Chem. Int. Ed., 54 (2015), pp. 13923–13927.

[82] L. Monico, K. Janssens, C. Miliani, G. Van der Snickt, B. Brunetti, M. Cestelli Guidi, M. Radepont, and M. Cotte, *Degradation process of lead chromate in paintings by Vincent van Gogh studied by means of spectromicroscopic methods. 4. artificial aging of model samples of co-precipitates of lead chromate and lead sulfate*, Analytical Chemistry, 85 (2013), pp. 860–867.

[83] L. Monico, K. Janssens, F. Vanmeert, M. Cotte, B. Brunetti, G. Van der Snickt, M. Leeuwestein, J. Salvant Plisson, M. Menu, and C. Miliani, *Degradation process of lead chromate in paintings by Vincent van Gogh studied by means of spectromicroscopic methods. part 5. effects of nonoriginal surface coatings into the nature and distribution of chromium and sulfur species in chrome yellow paints*, Analytical Chemistry, 86 (2014), pp. 10804–10811.

[84] L. Monico, G. Van der Snickt, K. Janssens, W. De Nolf, C. Miliani, J. Verbeeck, H. Tian, H. Tan, J. Dik, M. Radepont, and M. Cotte, *Degradation process of lead chromate in paintings by Vincent van*

Gogh studied by means of synchrotron X-ray spectromicroscopy and related methods. 1. artificially aged model samples, Analytical Chemistry, 83 (2011), pp. 1214–1223.

[85] J. Morris, *The Clinical Dementia Rating (CDR): Current version and scoring rules*, Neurology, 43 (1993), pp. 2412–2414.

[86] A. M. Mubeen, A. Asaei, A. H. Bachman, J. J. Sidtis, and B. A. Ardekani, *A six-month longitudinal evaluation significantly improves accuracy of predicting incipient alzheimer's disease in mild cognitive impairment*, Journal of Neuroradiology, 44 (2017), pp. 381–387.

[87] S. Mukherjee, S.-E. Choi, M. L. Lee, P. Scollard, E. H. Trittschuh, J. Mez, A. J. Saykin, L. E. Gibbons, A. F. T. M. A. Sanders, R. E.and Zaman, W. A. Kukull, D. A. L. A. Z. L. E. B. Barnes, L. L.and Bennett, M. Cuccaro, S. Mercado, L. Dumitrescu, T. J. Hohman, and P. K. Crane, *Cognitive domain harmonization and co-calibration in studies of older adults.*, Neuropsychology, (2022).

[88] A. Munoz, V. Carey, J. P. Schouten, M. Segal, and B. Rosner, *A parametric family of correlation structures for the analysis of longitudinal data*, Biometrics, 48 (1992), pp. 733–742.

[89] R. I. Pfeffer, T. T. Kurosaki, C. H. Harrah, Jr., J. M. Chance, and S. Filos, *Measurement of Functional Activities in Older Adults in the Community*, Journal of Gerontology, 37 (1982), pp. 323–329.

[90] G. Polatkan, S. Jafarpour, A. Brasoveanu, S. Hughes, and I. Daubechies, *Detection of forgery in paintings using supervised learning*, in 2009 16th IEEE International Conference on Image Processing (ICIP), 2009, pp. 2921–2924.

[91] A. Rathore and H. Winkle, *Quality by design for biopharmaceuticals*, Nat Biotechnol, 27 (2009), pp. 26–34.

[92] D. Riddle, *(Eds.) Brain Aging: Models, Methods, and Mechanisms (1st ed.)*, CRC Press/Taylor & Francis, 2007.

[93] M. J. Rovine and P. A. McDermott, *Latent growth curve and repeated measures anova contrasts: What the models are telling you*, Multivariate Behavioral Research, 53 (2018), pp. 90–101. PMID: 29220588.

[94] C. Sandoval, E. Pirogova, and M. Lech, *Two-stage deep learning approach to the classification of fine-art paintings*, IEEE Access, 7 (2019), pp. 41770–41781.

[95] M. D. Schwartz, *Lecture 17: color.*

[96] S. L. Simpson, L. J. Edwards, K. E. Muller, P. K. Sen, and M. A. Styner, *A linear exponent ar(1) family of correlation structures*, Statistics in Medicine, 29 (2010), pp. 1825–1838.

[97] L. M. Sullivan, *Repeated measures*, Circulation, 117 (2008), pp. 1238–1243.

[98] Teejet Technologies, *Technical Information.*

[99] M. Terashima, Y. Murai, M. Kawamura, S. Nakanishi, T. Stoltz, L. Chen, W. Drohan, R. L. Rodriguez, and S. Katoh, *Production of functional human α_1-antitrypsin by plant cell culture*, Applied Microbiology and Biotechnology, 52 (1999), pp. 516–523.

[100] K. van den Berg, E. Hendriks, M. Geldof, S. de Groot, I. van der Werf, C. Miliani, P. Moretti, L. Cartechini, L. Monico, M. Iwanicka, P. Targowski, M. Sylwestrzak, and W. Genuit, *Structure and Chemical Composition of the Surface Layers in the Amsterdam Sunflowers*, Amsterdam University Press, Amsterdam, Netherlands, 2019, pp. 159–173.

[101] N. van Noord, E. Hendriks, and E. Postma, *Toward discovery of the artist's style: learning to recognize artists by their artworks*, IEEE Signal Processing Magazine, 32 (2015), pp. 46–54.

[102] L. van Tilborgh, *Van Gogh and the Sunflowers*, Van Gogh Museum, Amsterdam, NL, 2008.

[103] L. van Tilborgh and E. Hendricks, *The Tokyo* Sunflowers*: a genuine repetition by Van Gogh or a Schuffenecker forgery?*, Van Gogh Museum Journal, (2001), p. 16.

[104] F. Vanmeert, E. Hendriks, G. Van der Snickt, L. Monico, J. Dik, and K. Janssens, *Chemical mapping by macroscopic X-ray powder diffraction (MA-XRPD) of Van Gogh's sunflowers: Identification of areas with higher degradation risk*, Angewandte Chemie (International ed. in English), 57 (2018), pp. 7418–7422.

[105] K. Weinfurt, *Repeated measures analysis: Anova, manova, and hlm*, Reading and Understanding More Multivariate Statistics, (2012).

[106] G. Wyszecki and W. S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae, 2nd Edition*, 2000.

[107] D. Xiao and J. Ohya, *Contrast enhancement of color images based on wavelet transform and human visual system*, in Proceedings of the IASTED International Conference on Graphics and Visualization in Engineering, GVE '07, USA, 2007, ACTA Press, p. 58–63.

[108] Y. Zhao, R. Berns, L. Taplin, and J. Coddington, *An investigation of multispectral imaging for the mapping of pigments in paintings*, in Computer Image Analysis in the Study of Art, D. G. Stork and J. Coddington, eds., vol. 6810, 2008, pp. 65–73.

[109] B. Zhou, S. Xu, and X.-x. Yang, *Computing the color complexity of images*, in 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2015, pp. 1898–1902.