# DATA QUALITY CONTROL AND INTER-FUNCTIONAL ANALYSIS ON DYNAMIC PHENOTYPE-ENVIRONMENTAL RELATIONSHIPS

By

Lei Xu

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science - Master of Science

2016

## ABSTRACT

DATA QUALITY CONTROL AND INTER-FUNCTIONAL ANALYSIS ON DYNAMIC PHENOTYPE-ENVIRONMENTAL RELATIONSHIPS

By

Lei Xu

Plant phenomics have become essential component of modern plant science. Such complex data sets are critical for understanding the mechanisms governing energy intake and storage in plants. Large-scale phenotyping techniques have been developed to conduct high-throughput phenotyping on plants. However, a major issue facing these efforts is the determination of the quality of phenotypic data. Automated methods are needed to identify and characterize alteractions caused by system errors, all of which are difficult to remove in the data collection step. Another issue is we are limited by the tools to analyze fully the phenomics data, *esp.* the dynamic relationships between environments and phenotypes.

The overarching goal of this thesis is to explore dynamic phenotype-environmental data via data mining/machine learning methods. Raw data measured from biological devices is pre-processed to numerical data, then cleaned by Dynamic Filter to ensure high data quality for further analysis. The cleaned data is further explored and applied with inter-functional analysis in order to find patterns that comply with both machine learning methodologies and biological constraints.

In this thesis we developed two tools to make exploration of phenotyping data available: (1) For data quality control, we developed a coarse-to-refined model called Dynamic Filter to identify abnormalities in plant photosynthesis phenotype data. (2) For inter-functional phenomics data analysis, we present a new algorithm called PhenoCurve for inter-functional phenomics data analysis.

## ACKNOWLEDGMENTS

First and foremost, I feel indebted to my advisor, Professor Jin Chen, for his guidance, encouragement, and inspiring supervision throughout the course of this research work. His patience, prudential attitude, extensive knowledge, and creative thinking have been the source of inspiration for me. He was available for advice or academic help whenever I needed and gently guided me for deeper understanding. When I hesitated between research program and course program two years ago, I was not sure about my decision, but now I am so happy and proud to say that I made a so wise decision. It is extremely hard to express how grateful I am for his unwavering support over the last two years and in the coming future.

I would like to take on this opportunity to thank my thesis committee members Professor Arun Ross and Professor Yanni Sun who have accommodated my timing constraints despite their full schedules, and provided me with precious feedback for the presentation of the results, in both written and oral form. I would also like to thank Professor Rong Jin for his help and advice during my graduate program studies.

During my Master studies, I had the pleasure of collaborating with many researchers from each and every one of which I had things to learn, and the quality of my research was considerably enhanced by these interactions. I would like to thank David M Kramer, Jefferey A Cruz, Linda J Savage, Qiaozi Gao, Oliver Tessmer, Yifan Yang, and Zheyun Feng for all the discussions we had and the fun moments we spent together on doing research and future planning.

Living in East Lansing without my good friends would not have been easy. I want to thank all my friends in the department and outside the department.

Last but definitely not least, I want to express my deepest gratitude to my beloved

parents and dearest girlfriend. Their love and unwavering support have been crucial to my success, and a constant source of comfort and counsel.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ALGORITHMS

# Chapter 1

# Introduction

Plants capture sunlight to fix $CO_2$ into energy rich molecules, thus supplying our ecosystem with $O_2$ and essentially all of its biological energy, including 100% of our food. In plants, photosynthesis is the primary energy source for metabolism and growth. Understanding how plants optimize or regulate it in response to a continuously changing and unpredictable environment is an essential component for developing strategies to improve crop yields to meet our growing needs for food and fuel [32]. Recent work has focused on improving the efficiency of photosynthesis to meet our growing needs for food and fuel ([9, 65, 32]). In order to develop efficiency-boosting mechanisms that reduce energy losses or enhance $CO_2$ delivery to cells during photosynthesis, advanced technologies in high-throughput plant photosynthetic phenotyping and pheno-informatics have been developed ([74], [29, 61, 14]). These technologies have allowed plant photosynthesis phenotypic variability to be characterized and to be related to putative biological functions, leading to a better understanding of the underlying mechanisms that control photosynthetic properties under various environmental conditions. Plant phenomics is a first-class asset for understanding the mechanisms regulating energy intake in plants ([52, 19]).

Plant phenotyping systems monitor photosynthetic performance for many plants both continuously and simultaneously. Phenomics datasets are large and continue to grow as we increase duration of sampling and resolution. Yet despite the size and richness of the data, small clusters of erroneous values, which give the appearance of real differences in

biological responses, can skew the analysis towards an invalid interpretation ([27]). There are several ways in which a measurement can be in error: errors originating from instrumentation malfunctions, biased values from miscalibrated sensors, and inevitable errors of precision. All these issues compromise the downstream data analysis tasks. Given the value of clean data for any operation, the ability to improve data quality is a key requirement for effective knowledge mining from large-scale phenotype data.

Advanced technologies in high-throughput plant photosynthetic phenotyping have been recently developed, allowing various photosynthesis parameters (such as $\Phi_{II}$, $q_E$, $q_L$, $q_I$, and $NPQ$) to be to characterized and to be related to putative biological functions [29, 52, 59, 14]. In parallel, machine learning and computer vision algorithms have been developed for phenotype information retrieval, data quality control, and knowledge discovery [24, 62, 71, 72, 69, 21].

With the large volume of plant phenotype data been generated, normalized and cleaned, biologists expect to immediately identify mutant strains with efficient photosynthesis machinery, and quickly generate and test biological hypotheses that may lead to new breakthrough in bio-energy research. Indeed, computational tools have been developed to identify temporal patterns from phenomics data [70], to group plants by phenotypes [21], and to predict unknown gene functions [64]. Nevertheless, it has been emphasized in literature that it is crucial to simultaneously measure and model multiple kinds of phenotypes and environmental factors to arrive at a holistic characterization of plant performance [68, 33, 25, 66]. Specifically, photosynthesis must respond to changing environment to provide the optimal amount of energy to meet the needs of the organism, in the correct forms, without producing toxic byproducts, e.g. reactive oxygen species or glycolate [32]. In this context, photosynthesis is a set of integrated modules that form a self-regulating network modulated by signal

transduction.

This chapter is devoted to an overview of these two broad topics of exploring phenotyping data, aiming to develop a general correspondence from data cleaning to inter-functional analysis. Here we move towards to the definitions in a fairly non-technical manner and the formal detailed definitions will be given in Chapter 2.

## 1.1 Data quality control

In this thesis, we focus on data abnormalities detection, which is a type of measurement error, in order to demonstrate how clean phenotype data can be obtained. Similar to sensor data, abnormalities in plant phenotype data deviate significantly from expected patterns and are visible outliers in the whole dataset ([56, 58]). The majority of abnormalities in plant phenotyping originate from instrumentation malfunctions (e.g., loss of sensor synchronization during measurement), or non-biological statistical outliers caused by data collection limitations (e.g., deterioration of signal-to-noise ratio for a sample as it progresses through the experiment).

Data abnormalities are often viewed as outliers in the whole dataset. Recent work has shown the effectiveness of applying data mining techniques, especially outlier detection, for the purpose of data cleaning ([39]), making it possible to automate the cleansing process for a variety of domains ([48, 12, 40, 17]). In these methods, by detecting the minorities of values that do not conform to the general characteristics of a given data collection, outliers are identified and are considered violations of association rules or other patterns in the data. However, the existing models are not suitable for phenotype data cleaning. These methods, while applied to phenotype data, may remove outliers including both measurement errors

and true biological discoveries, since true biological discoveries, to some extent, are outliers as well. Furthermore, detecting abnormalities from long time-series phenotype data requires handling a high temporal dimension, which increases the model complexity.

In order to identify and remove abnormalities in phenotype data and to minimize the deletion of biological discoveries, we have developed a coarse-to-refined residual analysis algorithm, called *Dynamic Filter*. Dynamic Filter has three key steps: 1) identify abnormal candidates at the coarse level, 2) refine abnormality identification in a projected feature space, and 3) iteratively identify abnormalities at the refined level. Dynamic Filter can speed up the data preparation process and make it more effective. Such improvements will minimize time consuming and labor intensive data preparation and increase the significance and confidence in biological discoveries. In summary, our model has the following advantages:

* To our knowledge, Dynamic Filter is the first work to integrate biological constraints with time-series phenotype data for data cleaning.

* Our model can identify both abnormalities and biological discoveries.

* Dynamic filter outperforms the existing solutions by optimizing the fitness between phenotype data and biological constraints.

## 1.2 Inter-functional Analysis

Studying the complex relationships between phenotype and environment poses several computational challenges. First, given that the phenotype-environment relationship is largely unknown, people tend to learn the inter-functional relationship using data driven approaches such as linear regression or curve fitting. However, 1) it is difficult to choose the best fitting

function since the relationship is usually non-linear [55]; 2) biological knowledge for describing the organism responses to environmental changes cannot be effectively incorporated into these models; 3) the purely data driven approaches may be significantly affected by outliers and bias in phenomics data, resulting in inaccurate phenotype-environment relationships. Second, if a phenotype-environment relationship has been well studied, people tend to directly fit data onto a known biological model (such as Michaelis-Menten kinetics, one of the best known models of enzyme kinetics) [16, 11]. Using a biological model, phenotype data can be converted into less complicated model parameter values, which are biologically meaningful and are easier to interpret than raw data [18]. However, biological models are simple compared with the real world situation. It is inappropriate to directly use a static theoretical model to learn dynamic relationships that vary over time and environmental conditions [69]. Finally, photosynthesis phenotypes are usually measured under dynamic environmental conditions, in a relatively long time period, and on many plants with vastly different genetic backgrounds. The broad range of data variation adds another level of complexity to the problem. In summary, new inter-functional algorithms are required to explore the complex phenotype-environment relationships.

In this article, we present a comprehensive data analysis approach called *PhenoCurve* to explore the dynamic phenotype-environment relationships. Comparing to the existing methods that solely model phenotypes or environmental factors, PhenoCurve, which enables researchers to model phenotypes and environmental factors simultaneously, has three major advantages. First, although phenotype and environment are measured separately with different sensor techniques, they are biologically correlated. Studying the inter-functional relationship may reveal patterns that cannot be discovered by only using phenotype or environmental data. Second, the inter-functional analysis allows us to condense the huge amount

of phenotype data into a succinct, highly compacted form, which will benefit all the followed data mining processes such as clustering and gene ranking. Third, different to purely data driven approaches, PhenoCurve is able to further improve its performance by incorporating precious biological knowledge on plant photosynthesis phenotypes. In the following content, we demonstrate the effectiveness of PhenoCurve by identifying the dynamic relationships between a key photosynthesis phenotype $\Phi_{II}$ (steady state quantum yield of photosystems II) and light intensity (denoted as $i$). PhenoCurve can be easily extended for other phenotype and environment data with simple modification.

## 1.3 Thesis Contributions

In this section we shall elaborate on the main problems considered in this thesis and our key contributions to address these problems.

This dissertation mainly deals with the exploring phenotyping data using machine learning/data mining techniques, including data quality control and inter-functional analysis. Generally, we attempt to develop an application that is able to identify outliers with high performance, and we attempt to design an algorithm for reliable and robust curve fitting for phenotyping values.

- **Dynamic Filter**. The thesis proposes a coarse-to-refined model called Dynamic Filter to identify abnormalities in plant photosynthesis phenotype data by comparing light responses of photosynthesis using a simplified kinetic model of photosynthesis. Dynamic Filter employs an Expectation-Maximization process to adjust the kinetic model in coarse and refined regions to identify both abnormalities and biological outliers. The experimental results show that our algorithm can effectively identify most

of the abnormalities in both the real and synthetic datasets.

- **Phenocurve fitting for inter-functional analysis**. The thesis presents a new algorithm called PhenoCurve for inter-functional phenomics data analysis. PhenoCurve is a model based curve fitting algorithm aiming to study both the values and the changing rates of the dynamic phenomics data. The evaluation on the real and simulated phenotype data showed that PhenoCurve has the best performance among all the tested methods. Its application on real plant photosynthesis data revealed new functions of core genes and processes that control photosynthesis efficiency in response to varying environmental conditions, which are critical for understanding plant energy storage and improving crop productivity.

## 1.4 Thesis Overview

The remainder of this dissertation is organized as follows. Chapter 2 lays out the foundation for the rest of the thesis.

The first part of the thesis focuses on data quality control problem. In Chapter 3 we focus on data quality control and the application Dynamic Filter developed , investigate how it is proposed and try to address real case problem of low data quality.

The second part of the thesis deals with inter-functional analysis problem. Chapter 4 discusses the motivation and context of doing inter-functional analysis for phenotyping data, especially after we obtained high data quality using Dynamic Filter.

Finally, Chapter 5 summarizes this work by concluding the main contributions, some potential extensions and the future research directions. In order to facilitate independent reading of various chapters, some of the definitions are repeated for multiple times.

## 1.5 Bibliographic Notes

Some of the results in this dissertation have appeared in prior publications. The material in Chapter 3 is based on a work published in the Bioinformatics on Oxford Journal [69] (Bioinformatics) and the content of Chapter 4 comes from current work which is to be submitted.

# Chapter 2

# Background

The goal of this chapter is to give a general and formal overview of the materials related to the work that has been done in this thesis. In particular, we will discuss the key concepts and questions relevant to problems of data cleaning and curve fitting. The exposition given here is necessarily very brief and the detailed discussion will be provided in the relevant chapters.

## 2.1 Data cleaning for high data quality

Data cleaning is the process of identifying incorrect or corrupted records in a dataset. The goal of data cleaning is to ensure an accurate representation of the real-world constructs to which the data refer. Removing impurities from data is traditionally an engineering problem, where ad-hoc tools made up of low-level rules (such as detecting syntax errors) and manually tuned algorithms are designed for specific tasks (such as the elimination of integrity constraints violations) ([44]). Detection and elimination of complex errors representing invalid values, however, go beyond the checking and enforcement of integrity constraints. They often involve relationships between two or more attributes that are very difficult to uncover and describe by integrity constraints. Recent work has shown the effectiveness of applying techniques from statistical learning for the purpose of data cleaning. In particular, outlier detection methods have made it possible to automate the cleansing process for a variety of

domains ([13, 17, 31, 39, 40, 48]).

### 2.1.1 Related work

None of the existing outlier-detection based methods are suitable for phenotype data cleaning. First, both biological discoveries and errors of detection are difficult to separate from distribution. Second, the cohesiveness rule used in temporal data cleaning is not applicable for the phenotype data, because 1) a non-cohesive time-serial could represent an interesting phenotype pattern rather than an error; 2) all the observations at the same time point may be similarly affected by a systematic abnormal event ([44]).

Alternatively, rather than checking the raw values, residue analysis can be employed to model the differences between the real values and the theoretical curve, which is usually derived from biological constraints such as the generalized light reactions ([30, 38]). This is often called the *goodness-of-fit* model. The goodness-of-fit based data cleaning models can be classified into two categories. First, statistical distribution characters such as mean, standard deviation, confidence interval or range have been used to find unexpected values indicating possible invalid values ([39]). Such simple methods can be efficiently applied to big data. However, these parameters (such as mean) are inclined to be biased by abnormalities with large deviations. Since it does not take into account local characteristics of data, there is a risk of mislabeling a range of normal data as abnormalities, and vice versa. Second, combined data mining techniques are used to identify patterns that apply to most residual records. A pattern is defined by a group of residuals that have similar characteristics (behavior for certain percentage of the fields in the dataset). Outliers are then identified as values that do not conform to the patterns in the data. Among them, the Hampel filter uses the median of neighboring observations as a reference value and looks for local outliers in a streaming

data sequence ([48, 49]). While the Hampel filter is suitable for temporal data cleaning, it assumes that the data are independent and identically distributed, which is not valid under dynamic environmental conditions. But when the data is highly autocorrelated, Hampel filter may fail to capture abnormalities.

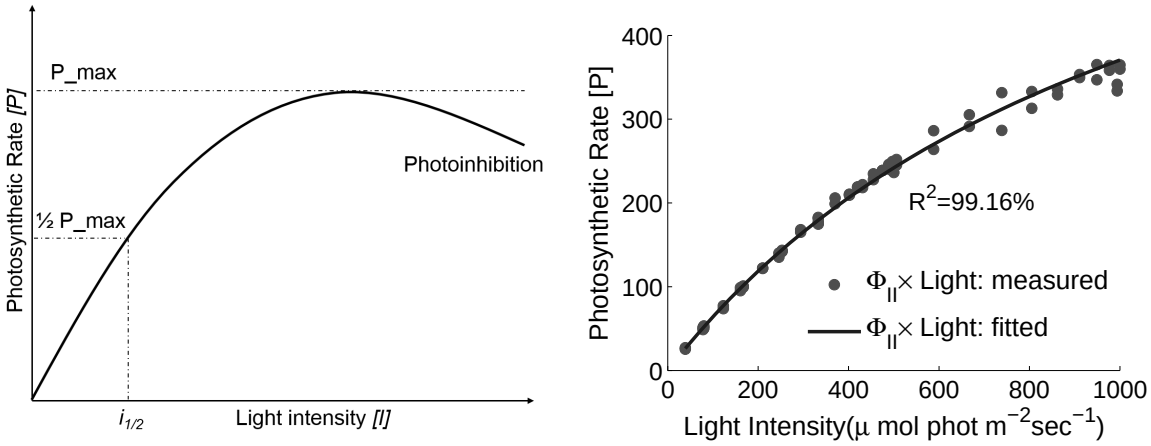
It should be noted that while the goodness-of-fit based data cleaning models focus on the modeling of deviation, they are not aware that the theoretical curve, which is used as the reference, may not always be precise. Typically, theoretical curves derived from biological knowledge are simple compared with the real world situation. It is therefore inappropriate to directly use the imperfect theoretical curve to infer abnormalities.

In this paper, we develop a coarse-to-refined residual analysis model called *Dynamic Filter* to effectively identify abnormalities in plant photosynthesis phenotype data. Our model derives a theoretical curve from the photosynthetic biological constraints; adjusts the theoretical curve to fit the phenotype data via optimization; and then studies the deviations of individual phenotype values from theoretical curve. The resulting patterns in residuals indicate abnormalities, which is a type of errors of detection, and the optimized theoretical curves reveal true biological outliers.

## 2.2 Inter-functional analysis

In this section, we introduce the biological background on modeling the relationship between light intensity and the rate of photosynthesis, as well as the existing computational approaches on curve fitting and regression.

The photosynthesis-irradiance (PI) curve is a graphical representation of the empirical relationship between solar irradiance and photosynthesis [38]. As a derivation of Michaelis-




(a) Hyperbolic Photosynthesis-Irradiance (PI) curve.

(b) Measured photosynthetic rates of Col-0 and the fitted PI curve.

Figure 2.1: Photosynthesis-Irradiance (PI) curve.

Menten kinetics, PI is modeled as a hyperbolic curve as shown in Fig. 2.1(a) [42], indicating that there is a generally positive correlation between light intensity and photosynthetic rate. The PI curve has been applied successfully to clean plant photosynthesis phenotype data, as well as to explain ocean-dwelling phytoplankton photosynthetic response to changes in light intensity [30, 69].

Let $P$ be photosynthetic rate at a given light intensity $[I]$, the PI function is in Eq. 2.1:

$$P = \frac{P_{max}[I]}{i_{1/2} + [I]} \tag{2.1}$$

where $P_{max}$ is the maximum potential photosynthetic rate per individual, and $i_{1/2}$ is half-saturation parameter representing the amount of light to product half of maximum photosynthesis. Parameter $i_{1/2}$ is usually learned with a serial of values of $\Phi_{II}$ and light $(i)$ using curve fitting methods.

By describing the photosynthetic rate $P$ using linear electron flow, we can associate both

$\Phi_{II}$ and $i$ with time $t$ using Eq. 2.1:

$$\Phi_{II}(t, i_{1/2}) = \frac{\max(\Phi_{II})}{1 + \frac{i(t)}{i_{1/2}}} \tag{2.2}$$

where $t$ is a time point in a user-defined temporal region $R$; $\Phi_{II}(t)$ and $i(t)$ represent $\Phi_{II}$ (steady state quantum yield of photosystems II) and light intensity at $t$; $\max(\Phi_{II})$ is the maximal $\Phi_{II}$ in $R$. See more details in [69].

During short time region, $i_{1/2}$ in Eq. 2.2 remains constant. However, it may change gradually over long time region with the changes of multiple environmental factors such as light, temperature, $CO_2$, and $O_2$ [14]. This phenomenon, namely phenotypic plasticity, indicates the ability of an organism to produce more than one phenotype when exposed to different environments [50, 46].

### 2.2.1 Related work

Knowing the trend of $i_{1/2}$ will greatly enhance our understanding to the mechanism that regulate plants in response to a continuously changing and unpredictable environment. However, to capture $i_{1/2}$ by directly applying the PI function on the phenotype and environment data may be inappropriate due to the high noise rate in real data [69].

#### 2.2.1.1 Sliding Window based Curve Fitting Methods

Since the change of $i_{1/2}$ is assumed to be smooth and continuous, a sliding window approach may be more appropriate than directly applying the PI function on every time point. Using the sliding window approach, we can divide the whole phenotype data into overlapped temporal regions, and then employ a curve fitting or regression method to compute a local

$i_{1/2}$ value for each region. Finally all the local $i_{1/2}$ values are merged to capture the global phenotype-environment relationship. Note that there is no explicit boundary between curve fitting and regression, while the former is more on the fitting optimization itself, and later focuses more on statistical inference [43].

Curve fitting is a commonly used method to model the relationships among two or more variables [43]. Mathematically, it is a process to tune the parameters of a known mathematical function $f(x)$ to achieve the best fit to a series of data points. The Levenberg-Marquardt algorithm (LMA), aka the damped least-squares method, has been widely used for nonlinear least squares calculations for solving generic curve-fitting problems [35]. LMA interpolates between the Gauss-Newton algorithm and the method of gradient descent, aiming to find a local minimum [28, 6].

If the fitting function $f(x)$ is unknown, kernel smoother, such as local linear regression (LLR), is often used for estimating a smooth function from a serials of noisy observations [26]. In this way, non-linear relationships between phenotypes and environmental factors can be learned from data, even if the underlying biological mechanism between them is unknown.

#### 2.2.1.2 Bayesian Linear Model with Normal Inverse Gamma Prior

While the sliding-window based curve fitting methods optimize every local fitting, they simply ignore the global continuity of the parameter $i_{1/2}$. Thus, they may be sensitive to noise and bias in local regions in raw phenotype data, resulting in inaccurate phenotype-environment relationships. To this end, performance improvement may be achieved by using Bayesian linear model with normal inverse gamma (NIG) prior [22].

First, we transfer Eq. 2.2 to its linear form for easier representation:

$$\frac{\max(\Phi_{II})}{\Phi_{II}(t_i)} - 1 = \frac{1}{i_{1/2}} i(t_i) + \epsilon(t_i) \tag{2.3}$$

where $\epsilon(t_i)$ is the error associated with $t_i$, distributed as normal distribution $N(0, \sigma_i^2)$.

Second, similar to Section 2.2.1.1, we adopt a sliding window approach to estimate $i_{1/2}$ and $\sigma_i^2$ for each temporal region using linear regression methods [20]. Given a threshold on the linear regression reliability $R^2$, we can classify all the temporal regions $D$ into two groups, i.e. reliable data $D_r$ and unreliable data $D_u$ [28].

Third, in order to derive the posterior probability $p(i_{1/2}, \delta^2 | \Phi_{II}, i)$, we assume that $(i_{1/2}, \sigma_i^2)$ follows the normal inverse gamma distribution, i.e. $i_{1/2}, \sigma_i^2 \sim NIG(\mu, V, a, b)$, where $\mu, V, a, b$ is the set of hyper parameters of NIG. Subsequently $i_{1/2}, \delta^2 | \Phi_{II}, i$ is also distributed as NIG yet with different hyper parameters $(\mu^*, V^*, a^*, b^*)$. Assuming that $D_r$ and $D_u$ follow the single mode of the NIG distribution, we estimate the hyper parameters using $D_r$ and apply them on $D_u$. Specifically, $i_{1/2}$ of each temporal region in $D_u$ can be obtained by 1) obtaining the marginal posterior distribution by integrating out $\sigma_i^2$, 2) maximizing a posteriori probability of the marginal posterior distribution, and 3) optimizing a Bayesian linear model using both the shared hyper parameters and local data. See more details at [22] and [8].

In this model, the global continuity of $i_{1/2}$ is guaranteed due to a single mode of the normal inverse gamma distribution (i.e. $D_r$ and $D_u$ share the same set of common hyper parameters) [8]. However, the assumptions on constant hyper parameters may be too rigid for real phenotype and environment data, in that $D_r$ and $D_u$ may be under different prior distributions due to different environmental conditions.

# Chapter 3

# Data cleaning with Dynamic Filter

A data cleaning framework called Dynamic Filter is proposed for data quality control in this Chapter.

The remainder of the chapter is organized as follows. Section 3.1 motivates the problem and main intuition behind the proposed algorithm, the detailed description of the proposed algorithm. Section 3.2 presents the detailed implementation issues, the experimental setup, results and analysis.

## 3.1 Methods

In this section, we first introduce the theoretical curve of time-series steady-state quantum yield data, and then introduce a framework for abnormality detection.

In this paper, the time-series steady state quantum yield of photo-systems II (denoted as $\Phi_{II}$) is chosen for abnormality detection for three reasons. First, $\Phi_{II}$ can be readily measured using fluorescence video imaging making it useful for high throughput phenotyping. Second, because it reflects light-driven electron transfer, it can be used as an indicator of photosynthetic rates and efficiency, albeit with the caveat that it reflects the sum of $CO_2$ fixation, photo-respiration and other processes ([47]). Finally, $\Phi_{II}$ is a good demonstration of the approach because it tends to follow, to a reasonable degree, relatively simple saturation behaviors. Given an adequate model, the cleaning procedure described in the manuscript

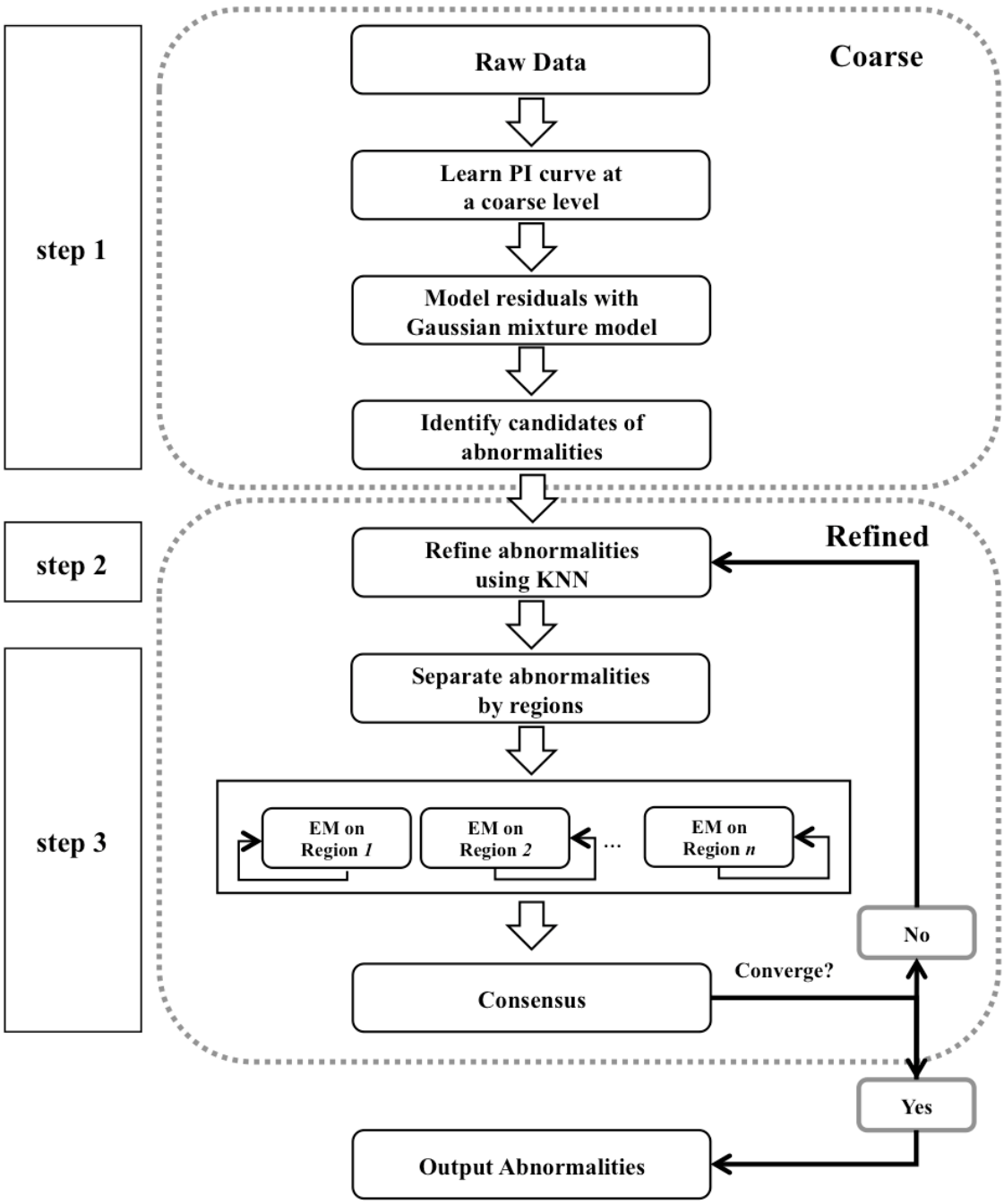



Figure 3.1: The framework of Dynamic Filter.

may also be applied to other photosynthetic parameters like non-photochemical quenching (NPQ), which can display complex behaviors.

### 3.1.1 Theoretical Photosynthetic Curve

An abnormality in residual analysis is an observation exhibiting a large difference between the theoretical value and the observed value, and may indicate a data entry error from the phenotyping sensors. To derive the theoretical curve, we model $\Phi_{II}$ with the photosynthesis-irradiance (PI) curve (see Fig 2.1(a)) ([30, 38]).

As a derivation of Michaelis-Menten kinetics, one of the best-known models of enzyme kinetics in biochemistry ([42]), PI is modeled as a hyperbolic curve (see Fig 2.1(a)) in Eq. 2.1, revealing the empirical relationship between solar irradiance and photosynthesis ([38]). where $P$ is photosynthetic rate at a given light intensity, $P_{max}$ is the maximum potential photosynthetic rate per individual, $[I]$ is a given light intensity, and $i_{1/2}$ is half-saturation constant. Fig 2.1(a) shows the generally positive correlation between light intensity and photosynthetic rate. The PI curve has already been applied successfully to explain ocean-dwelling phytoplankton photosynthetic response to changes in light intensity ([30]), as well as terrestrial and marine reactions.

We describe the photosynthetic rate $P$ in terms of linear electron flow ([32]), and associate both temporal steady state quantum yield of photosystems II $\Phi_{II}$ and temporal light intensity $i$ with time $t$, as shown in Eq. 2.2: where $t$ is a time point in a user-defined temporal region $T$ ($t \in T$); $\Phi_{II}(t)$ and $i(t)$ represent the steady state quantum yield of photosystems II and light intensity at $t$; $\max(\Phi_{II})$ is the maximal $\Phi_{II}$ in $T$; and the half-saturation constant $i_{1/2}$ is the light intensity at which the photosynthetic rate proceeds at half $P_{max}$. See proof in supplementary Section 1.

One may reasonably ask if the NPQ or photoinhibition would affect the theoretical model for light saturation. In fact, NPQ has (surprisingly) little effect on the relationship between

$\Phi_{II}$ and light intensity as can be readily seen in the fact that the $\Phi_{II}$ light saturation curves for wild type and the *npq*4 mutant of Arabidopsis are essentially identical despite large differences in qE (i.e., rapidly reversible photoprotection of NPQ) ([36]). The reason for this apparent disconnect is that, at high light, the slowest step in the light reactions of photosynthesis occurs subsequent to light absorption at the cytochrome $b6f$ complex and is finely regulated by the pH of the lumen ([60]). Light absorption become rate limiting only at NPQ levels much higher than those observed here. The biological role of NPQ under most conditions appears to be in regulating electron transfer but in preventing the buildup of reactive intermediates within the photosystem II reaction center ([45]). Thus, the effects of moderate levels of NPQ and photoinhibition should have little affect the behavior of the wild type system. However, under extreme conditions of in mutant lines with altered behavior producing high levels of NPQ or photoinhibition we expect to see behavior that deviates from that produced by the model. These instances will be detected as outliers and flagged for further investigation of possible biological discoveries.

Consequently, the half-saturation constant $i_{1/2}$ can be learned using all $\Phi_{II}(t)$ and $i(t)$ in $T$ with a nonlinear regression method ([54]). Note that the half-saturation constant can be dramatically different between plants and between leaves in plants. Thus, the general shape of the curve is typically maintained, but not its maximal or half-saturation light intensity.

Finally, given $i_{1/2}$, the residual value at each time point $t$ is defined as:

$$rsd(t) = \Phi_{II}(t) - \Phi'_{II}(t, i_{1/2}) \tag{3.1}$$

where $rsd(t)$ is the residual value at at time $t$; and $\Phi_{II}(t)$ is the observed value and $\Phi'_{II}(t)$ is the theoretical value of steady state quantum yield at $t$ calculated using Eq. 2.2.

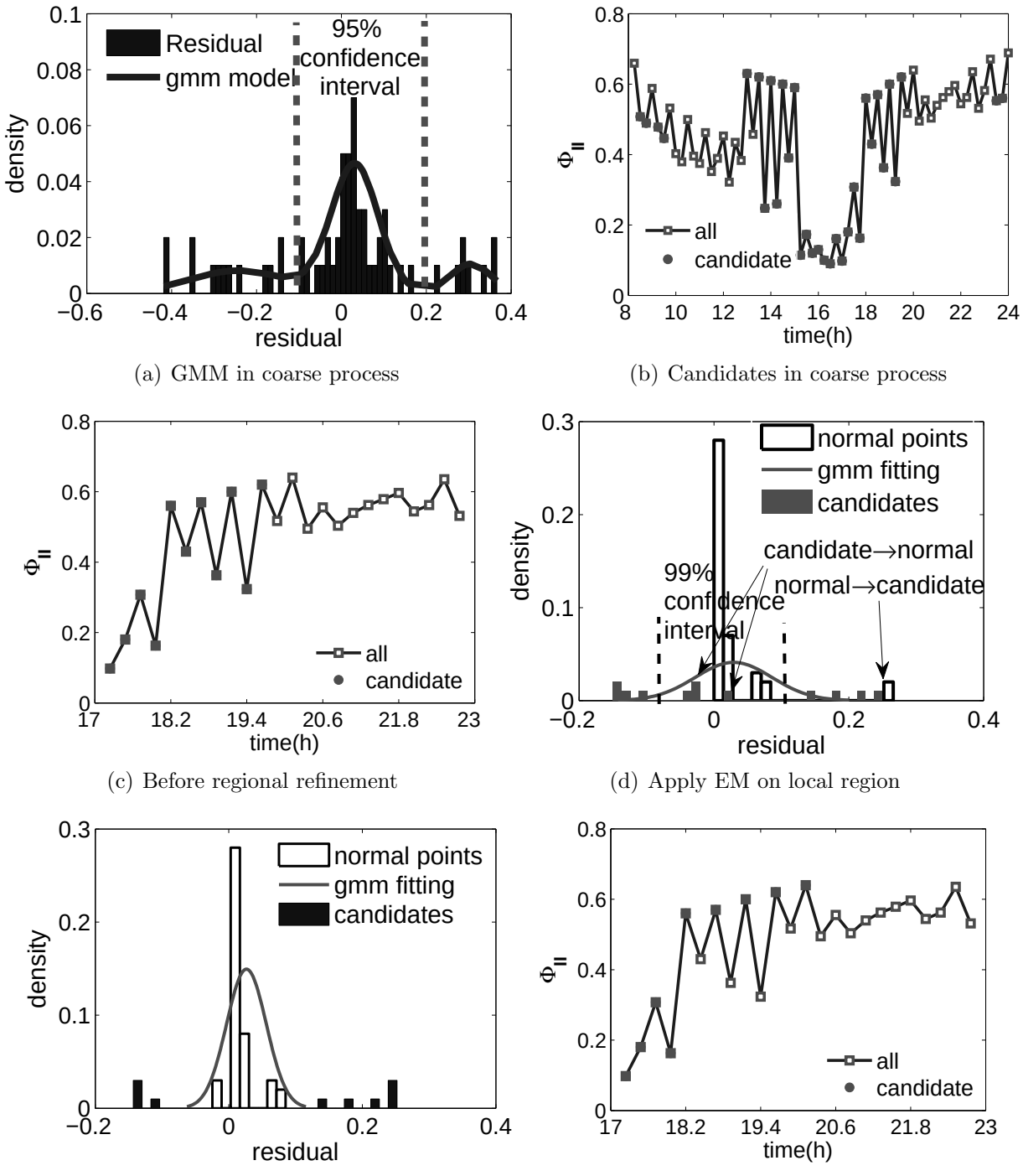
We note that there are multiple models for PI curves, which give similar responses to light ([37, 23, 73, 34, 15]). In this paper, we chose the Michaelis-Menten Kinetics model because it is convenient to use, and fits plant photosynthesis rate data well (see Fig 2.1(b)). It should be noted that an important feature of our approach is that these alternative models can be easily added or substituted for comparison.

### 3.1.2 Framework of Dynamic Filter

Dynamic Filter is a coarse-to-refined residual analysis approach, which has three major steps as shown in Fig. 3.1. We define abnormalities using a definition to that proposed in [40]:

**Definition 3.1. Abnormality.** *Let* $\{\Phi_{nor}\}$ *be a set of normal phenotype data, and* $\{rsd_{nor}\}$ *be the corresponding residual set. An abnormality* $\Phi_{abn}$ *is a phenotype value whose residual falling off the* $\alpha$ *confidence interval of the major normal distribution of* $\{rsd_{nor}\}$.

Note that confidence interval $\alpha = 99\%$ is commonly used in literature ([40, 57]), but is adjustable by users. In this paper, by adopting the concept of confidence interval, we assume that 1) the majority of the phenotype values are correct, and 2) they form the major distribution in the residual data, which is also distinctly different from the distribution(s) of the residual data of the abnormalities.




(a) GMM in coarse process

(b) Candidates in coarse process

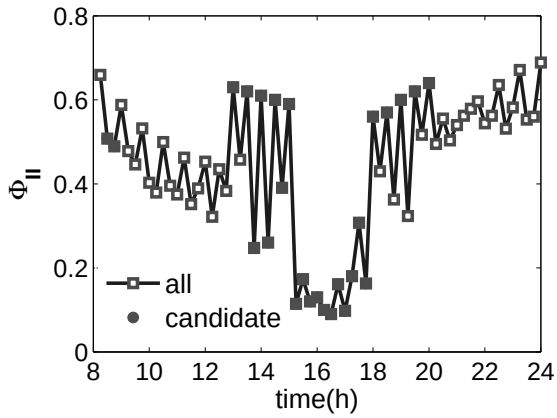
(c) Before regional refinement

(d) Apply EM on local region

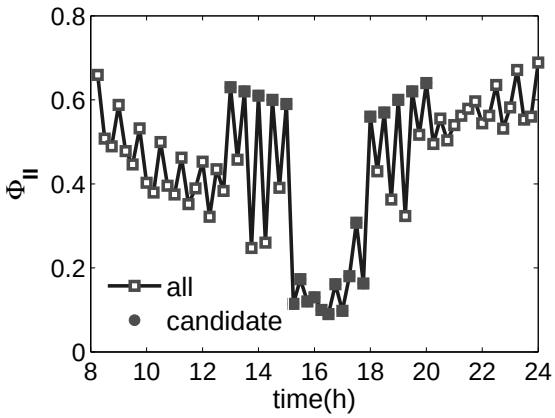
(e) Output of EM

(f) After regional refinement

Figure 3.2: An example of Dynamic Filter. The solid points are abnormalities and the hollow points are normal values.




(a) Candidates in refined process




(b) Final outputs

Figure 3.2 (cont'd)

#### 3.1.2.1 Step 1. Coarse process to identify abnormal candidates.

Given a set of phenotype data $\Phi_{II}$, we adopt Eq. 2.2 to generate the theoretical values of steady state quantum yields for each plant, denoted as $\{\Phi'_{II}\}$, by using the whole time-serial as temporal region $T$, aka the coarse level. For the dataset used in the Experiment Section, the smallest value of time interval is 10 minutes, and the scale of $T$ in the whole dataset is 3 days. Consequently, we generate the residual data of all plants $\{rsd\}$ using Eq. 3.1, and model them using a Gaussian mixture model (GMM) (see details in Section 3.1.3 and example in Fig. 3.2(a)). Finally, we generate the abnormality candidate set $\{\Phi_{abn}\}$ with Definition 3.1. In Fig. 3.2(b), solid points are abnormality candidates in the coarse process. Clearly, because of the simplified PI curve model, not all the abnormality candidates are correctly identified.

#### 3.1.2.2 Step 2. KNN process to refine abnormality identification.

Abnormality candidates may have certain intrinsic patterns of distribution highly related to certain ranges of feature space. For example, accidental dysfunction of data-capturing devices may cause abnormalities concentrated around some regions, which form statistical patterns on the distribution plot of the feature space. From a statistical viewpoint, abnormalities should be away from normal values in the feature space, and values with similar features tend to have the same labels. This leads to a refinement process to exploit the patterns of abnormalities candidates on selected feature space, and to make use of these patterns to refine abnormality identification, as described in Algorithm 1.

Specifically, we first select the optimal features from $\Phi_{II}$, $rsd$, $i$, $t$, etc., in which abnormalities and normal values are maximally separated. To solve this feature reduction problem, Linear Discriminant Analysis (LDA) is adopted to get the principal components

of the optimal feature space (Algorithm 2, see details in Section 3.1.3). Second, we apply K-nearest-neighbor approach on the selected feature space, such that each abnormality candidate will be relabeled as its majority label of $k$ nearest neighbors (Algorithm 1) ([4]).

**Algorithm 1** KNN process to refine results

**Input:**

- $\Psi$ : original feature space
- $C$: the set of labels (abnormality or normal)
- $k$: number of nearest neighbors used

$\Psi_{proj} \leftarrow$ Feature Selection$(\Psi, C)$ 2
**for** $\psi_i$ in $\Psi_{proj}$ **do**
    $C_i \leftarrow$ majority label of $k$ nearest neighbors
**end for**
**Output:** $C$

**Algorithm 2** Feature Selection

**Input:**

- $\Psi$ : original feature space
- $C$: the set of labels (abnormality or normal)

$\mu \leftarrow \frac{1}{|C|} \sum \psi$
**for** $i$ from 1 to 2 **do**
    $\Psi_i \leftarrow$ Features of $i_{th}$ label
    $n_i \leftarrow |\Psi_i|;\ \mu_i \leftarrow \frac{1}{n_i} \sum \psi_i$
    $SW_i \leftarrow \frac{1}{n_i} \sum (\psi_i - \mu_i)(\psi_i - \mu_i)^T$
**end for**
$SW \leftarrow \sum_{i=1}^{2} SW_i$
$SB \leftarrow \sum_{i=1}^{2} \frac{n_i}{|C|} (\mu_i - \mu)(\mu_i - \mu)^T$
$\Psi_{proj} \leftarrow eig(SB/SW) \cdot \Psi$
**Output:** $\Psi_{proj}$ : $\Psi_{proj}$ is projected space

**Step 3. Refined process to identify abnormalities in local regions.**

Because the theoretical values $\{\Phi'_{II}\}$ are learned with the simplified PI curve model at the coarse level, not all the assignments of the abnormal candidates are correct. Consequently, we separate the abnormal candidates $\{\Phi_{abn}\}$ to temporal checking regions (see Definition 3.2) and refine abnormality identification in each region.

**Algorithm 3** EM optimization on each local region $r$

**Input:**

- $\Phi$: phenotype values in a local region
- $i$: light intensity
- $\alpha$: confidence interval

1: Let $\Phi_{nor}$ and $\Phi_{abn}$ be normal values and abnormalities in $\Phi$
2: **repeat**
3: **E-step:**
4: $[rsd_{nor}, i_{1/2_{nor}}] \leftarrow$ PI_CurveFitting$(\Phi_{nor}, i)$
5: $[\mu_{nor}, \sigma_{nor}] \leftarrow GMM(rsd_{nor})$
6: $[rsd_{min}, rsd_{max}] \leftarrow$ getConfidenceInterval$(\mu_{nor}, \sigma_{nor}, \alpha)$
7: **M-step:**
8: $rsd_{abn} \leftarrow$ getResidual$(\Phi_{abn}, i_{1/2_{nor}})$
9: $[\Phi_{nor}, \Phi_{abn}] \leftarrow$ UpdateCandidate$(rsd_{nor}, rsd_{abn}, rsd_{min}, rsd_{max})$
10: **until** $\Phi_{nor}$ and $\Phi_{abn}$ are stable
11: **Output:** $\Phi_{nor}$

**Definition 3.2. Temporal Checking Region.** *A checking region $r$ consists of at most $m$ normal values flanking the selected abnormal candidates, depending on data availability, denoted as $\{\Phi_{nor}\}$, and at most $n$ abnormal candidates such that the last abnormal candidate is constrained to be at most $l$-timepoints away from the first one, denoted as $\{\Phi_{abn}\}$.*

In Definition 3.2, $m$, $n$ and $l$ are user defined parameters that determine the size of a temporal checking region. A check region has at most $m + l - n$ normal values and at most $n$ abnormities. Note that abnormal candidates can be continuous or discontinuous, and two checking regions may share common normal values.

In the refined process, an Expectation-Maximization (EM) process is employed to repeatedly optimize the results in each temporal region $r$. Pseudo-code of the EM process is shown in Algorithm 3. In the E step, using the local normal values $\{\Phi_{nor}\}$ in checking region $r$ as inputs, we regenerate the theoretical values $\{\Phi'\}$ with Eq. 2.2. Then the residuals $\{rsd\}$ for both the abnormal candidates $\{\Phi_{abn}\}$ and the normal values $\{\Phi_{nor}\}$ are regenerated using Eq. 3.1 (Algorithm 3 line 6-8). In the M step, we redefine the abnormal candidate set $\{\Phi_{abn}\}$

with the statistical distribution of the new residual data $\{rsd\}$ according to Definition 3.1. Specifically, a value falls off the confidence interval threshold of the major distribution of the normal residual values will be moved to $\{\Phi_{abn}\}$; and if an abnormal candidate is within the confidence interval threshold of the major distribution of the normal residual values, it will be labeled as normal and be moved to $\{\Phi_{nor}\}$ (Algorithm 3 line 10-11).

The EM process will stop when the label assignment is stable. Fig. 3.2(c-f) shows the iterative process in a checking region. Since checking regions may share common values, the results from different regions may be conflicted. For example, a phenotype value is identified as an abnormality in one region but is considered a normal value in another region. To solve conflicts and consequently improve performance, we employ an information sharing process in the end of the EM process to broadcast all the local results to all the checking regions. If conflict exists, voting results will be used to redefine abnormal candidates in the selected feature space (step 2), and the EM process will rerun on the new checking regions. The process will repeat till the results converge. Fig. 3.2(g,h) demonstrate that all the abnormalities are identified.

### 3.1.3 Related Works

We introduce the Gaussian Mixture Model (GMM) and the Linear discriminant analysis (LDA) used in Section 3.1.2.1 as follows.

#### 3.1.3.1 Gaussian Mixture Model.

A Gaussian Mixture Model is a parametric probability density function represented as a weighted sum of Gaussian component densities ([53]). GMMs are commonly used as a parametric model of the probability distribution of continuous features ([53]). The probability

density function is given by the equation:

$$p(x|\lambda) = \sum_{i=1}^{M} \omega_i g(x|\mu_i, \Sigma_i) \tag{3.2}$$

where $x$ is a D-dimensional continuous-valued vector, $\omega_i$, $i = 1, \ldots, M$, are the mixture weights, and $g(x|\mu_i, \Sigma_i), i = 1, \ldots, M$ are the component Gaussian densities. Each component density is a D-variate Gaussian function of the form:

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i)\} \tag{3.3}$$

with $\mu_i$ be the mean vector and $\Sigma_i$ be the covariance matrix ($i = 1, \ldots, M$). The mixture weights satisfy the constraint that $\sum_{i=1}^{M} \omega_i = 1$. GMM parameters are estimated from training data using the Maximum Likelihood Parameter Estimation or Maximum A Posteriori Estimation ([53]). In this paper, residuals are one-dimensional scalar data, we use $\mu_i$ and $\sigma_i$ to represent the mean and variance of residuals.

#### 3.1.3.2 Linear Discriminant Analysis for Feature Selection.

Linear discriminant analysis (LDA) is a methods used in statistics, pattern recognition and machine learning to find a linear combination of features, which characterizes or separates two or more classes of objects or events, such that the inter-class variance is maximized and the intra-class variance is minimized ([67]). The resulting combination may be used as a linear classifier, or more commonly, for dimensionality reduction before later classification. In this paper, we seek combination of features, with which normal values (one class) are centered around one area, while abnormalities (another class) are centered around a distinctively separated area.

Suppose there are $C$ classes, and each class has $n_i$ points, mean $\mu_i$ and intra-class variance $\Sigma_i$. Then the inter-class variance may be defined by the sample covariance of the class means:

$$SB = \sum_{i=1}^{C} \frac{n_i}{|C|}(\mu_i - \mu)(\mu_i - \mu)^T \tag{3.4}$$

and the intra-class variance of whole dataset is $SW = \sum_{i=1}^{C} SW_i$ ([41]). The class separation in a direction $\tilde{\omega}$ in this case will be given by:

$$S = \frac{\tilde{\omega}^T SB \tilde{\omega}^T}{\tilde{\omega}^T SW \tilde{\omega}^T} \tag{3.5}$$

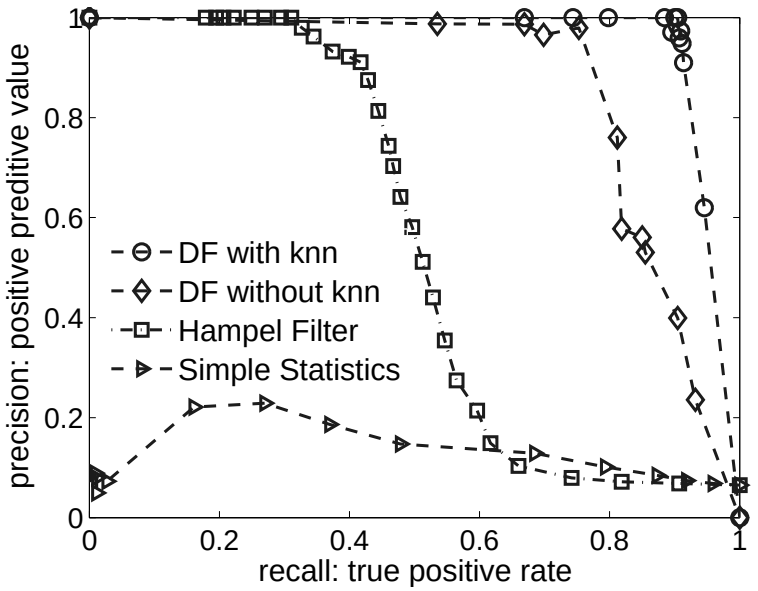
The objective function is to maximize $S$ and it can be shown that when $\tilde{\omega}$ is the eigenvector of $SW^{-1}SB$, $S$ will have maximized value corresponding to eigenvalue ([51]).

## 3.2 Experiment

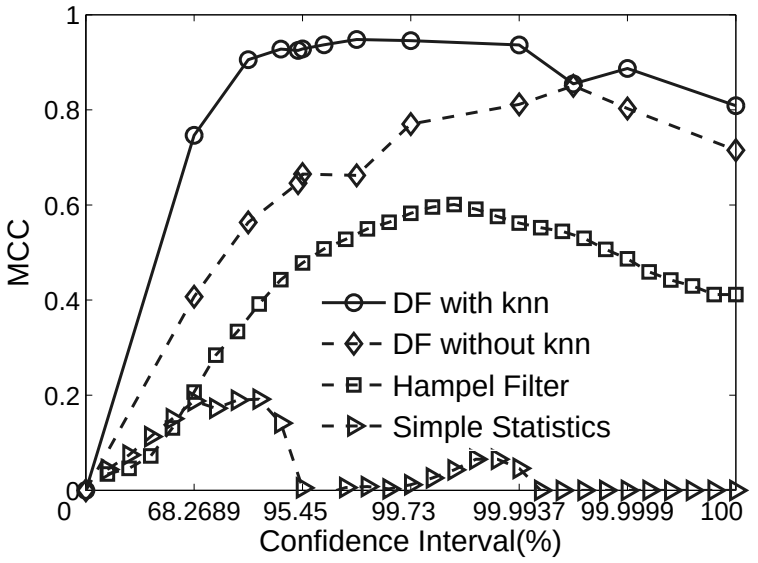
We compared Dynamic Filter on both real and synthetic datasets with two widely-used data cleaning algorithms: 1) a statistical approach that classifies abnormalities based on standard variance ([39]), and 2) Hampel filter that identifies abnormalities based on digress from median of trends ([48, 49]). Note that all the three methods were applied on the same phenotype residual data for a fair comparison.

For performance evaluation, we used both the precision-recall curve and the Matthews correlation coefficient (MCC) ([5]). The MCC that can appropriately represent a confusion matrix is computed with:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.6}$$




(c) Precision-Recall curves




(d) MCC w.r.t. confidence interval threshold

Figure 3.3: Performance evaluation of precision-recall and Matthews Correlation Coefficient on real dataset. DF represents Dynamic Filter.

### 3.2.1 Real phenotype dataset

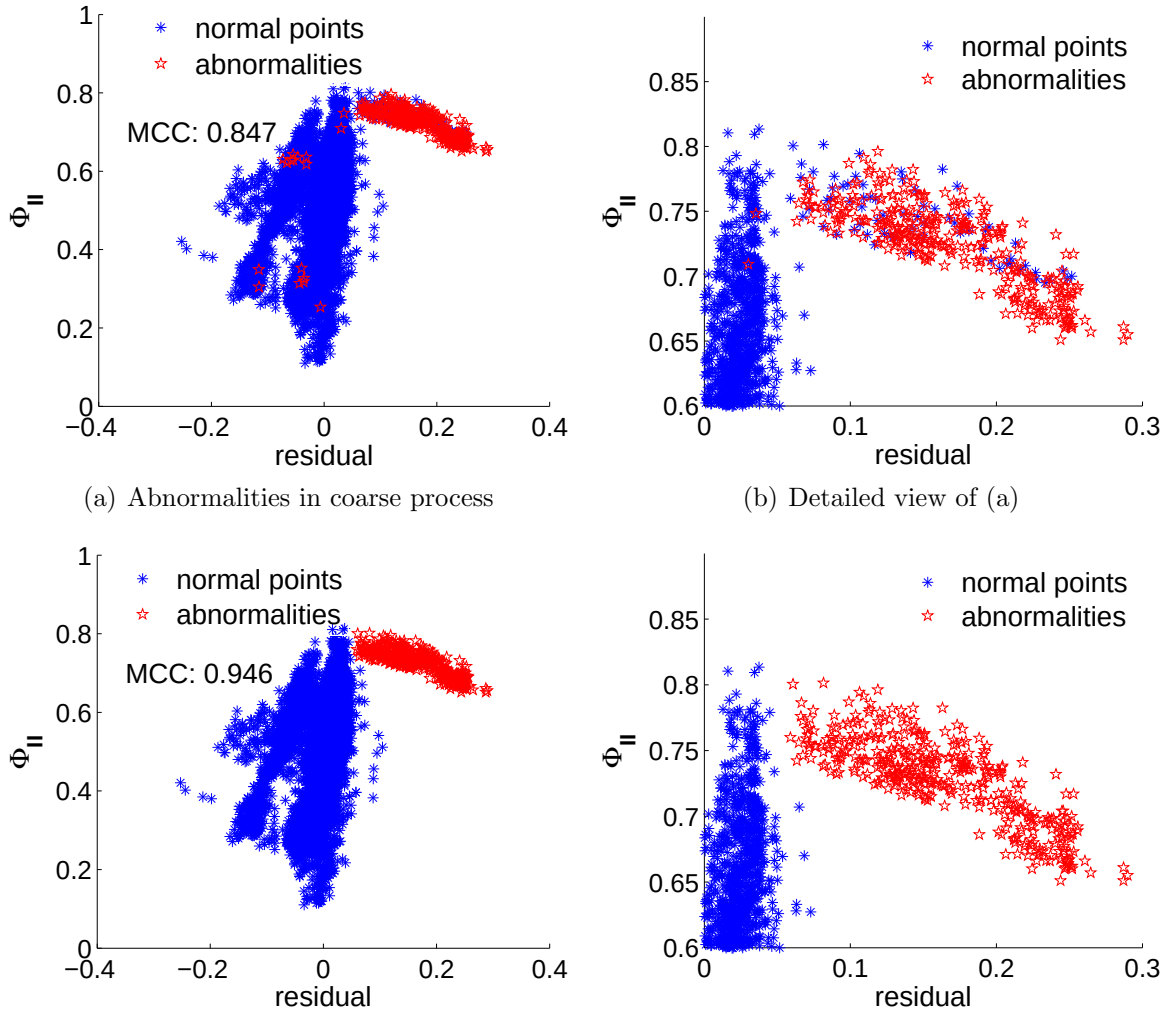
We first tested the performance of Dynamic Filter using the plant photosynthetic phenotype data consisting of 106 *Arabidopsis thaliana* plants (confirmed T-DNA insertion mutants and wild-types) sampled at 64 time points under dynamic light conditions ([3, 1]). The photo-

synthetic phenotype values vary dramatically across plants, reflecting potential differences in development, stress responses, or regulation of processes such as stomatal conductance, photodamage, and storage of photosynthate ([32]). Experts went through the data and manually marked the ground truth of abnormalities, and found the error rate is 6.5%.

The experimental results shown in Fig 3.3(a) indicated that Dynamic Filter is significantly better than the other two approaches in the precision-recall curve. Specifically, Dynamic Filter yields AUC as high as 0.964, higher than the AUC of simple statistics and Hampel Filter (0.147 and 0.543 respectively). Fig 3.3(b) shows our model is also significantly better according to MCC. Furthermore, it shows that Dynamic Filter is insensitive to the selection of the confidence interval threshold, which is distinctly different from the other algorithms that rely on well-picked parameters.

Note that the AUC of Dynamic Filter without KNN is 0.862 (Fig 3.3(a)), implying the KNN refinement (step 2) is a key component of Dynamic Filter. Specifically, Fig 3.4 shows how the KNN refinement improved the performance of data cleansing. On the $\Phi_{II}$ vs. residual plot shown in Fig 3.4(a) (detailed visualization on Fig 3.4(b)), some isolated normal values are misclassified as abnormalities, and certain abnormalities misclassified as normal values. Clearly, these values do not conform with the most nearby values. By applying the KNN refinement, these misclassification are effectively corrected (Fig 3.4(c,d)).

We systematically tested the performance of the different components of Dynamic Filter. Fig. 3.5 shows the performance improvement by comparing Dynamic Filter with a model without KNN refinement (v5), iteration of EM (v4), consensus on all regions (v3), reassignment of normal values and abnormalities in EM (v2), or even without the whole refined process (v1). It implies that the refined process, especially the KNN and EM refinement, is the key of performance improvement.




(a) Abnormalities in coarse process

(b) Detailed view of (a)

(c) Abnormalities after KNN

(d) Detailed view of (c)

Figure 3.4: Performance improvement by applying the KNN refinement process.

Fig 3.6 and 3.7 show case studies on the real data. In Fig 3.6, the experiment was run on a wild-type reference plant, Arabidopsis Col-0. In the coarse process, the residual analysis was applied to identify the abnormal candidates (Fig 3.6(a) and solid points in Fig 3.6(b)). Clearly, 6 solid points on the bottom were incorrectly labeled as abnormalities, which were gradually corrected in the refined process (Fig 3.6(c,d)). Fig 3.7(a) shows a true biological discovery on the real data. Our screen revealed accession ELY exhibiting photosynthetic characteristics markedly different from the reference (Col-0). It would, however, be labeled as abnormal and subsequently deleted by the existing outlier detection based data cleaning

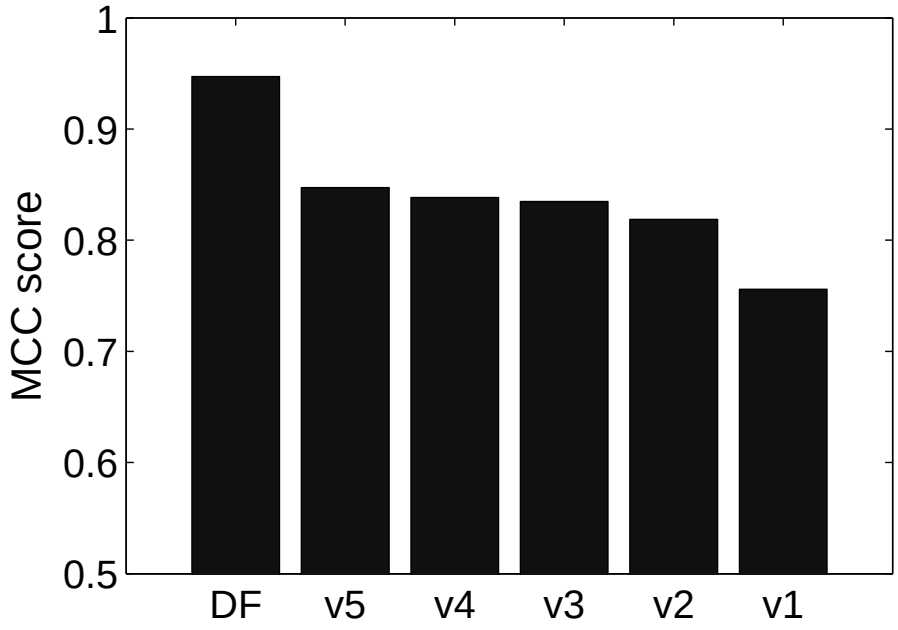
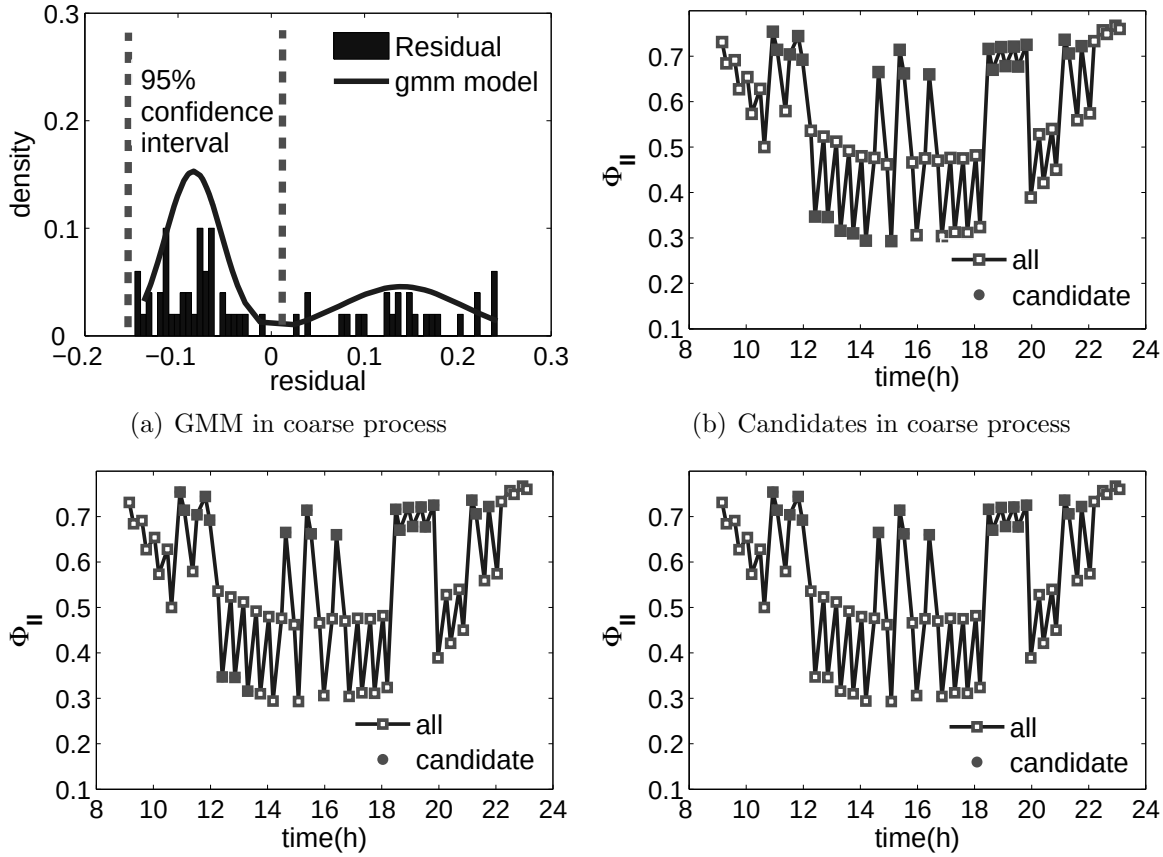



Figure 3.5: Performance comparison. Each version corresponds to a different version of Dynamic Filter (DF) without: KNN refinement (v5), iteration of EM (v4), consensus on regions (v3), reassignment of normal/abnormal labels in EM (v2), or the whole refined process (v1).

methods, resulting in over-clean problem. Dynamic Filter identifies ELY correctly and suggests that the differences in its quantum yield is caused by the monotone decrease of $i_{1/2}$ regardless the change of sunlight (see Fig 3.7(b)). The nonnegligible deviation between the observed values and the theoretical curve learned from the coarse phase of Dynamic Filter (see Fig 3.7(c)) implies the theoretical model is simple compared with the real world situation. Instead of directly use the PI curve to infer abnormalities, we optimize the fitting results in the refined phase of Dynamic Filter, resulting in almost perfect match between the observed values and the theoretical curve (see Fig 3.7(d)).

Furthermore, we varied the size of the temporal checking region and compared the performance in Fig 3.8. The results in Fig 3.8(a) reveal that Dynamic Filter achieves the best performance when $m$ is between 10 and 15. This number allows enough training data for the refinement process, meanwhile avoiding NPQ variation over long time interval. Fig 3.8(b)




(a) GMM in coarse process

(b) Candidates in coarse process

(c) Candidates in refined process

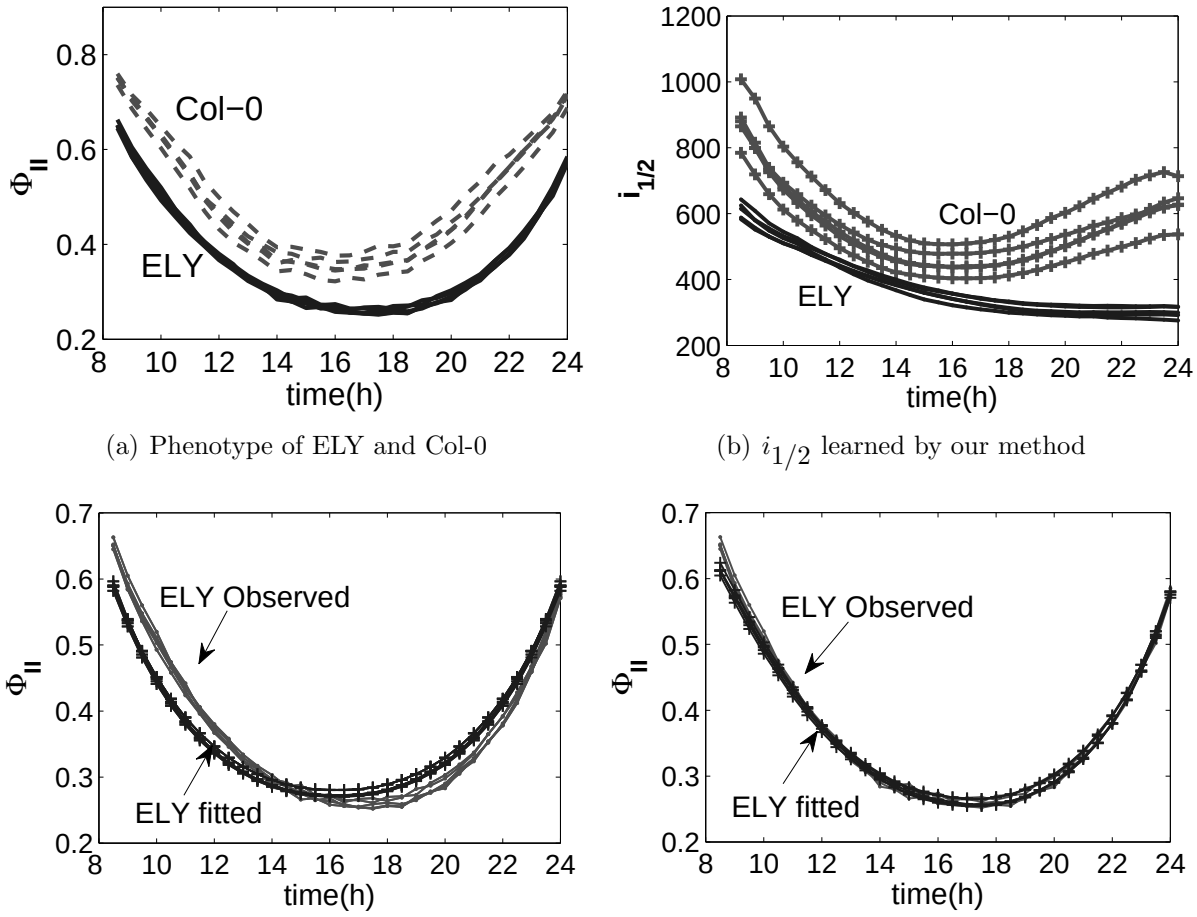
(d) Final outputs

Figure 3.6: A case study on the real data shows that Dynamic Filter correctly identifies all the abnormalities.

shows performance of Dynamic Filter is relatively stable against max number of abnormalities $n$, implying the robustness of Dynamic Filter is high.

### 3.2.2 Synthetic dataset

Since the true biological discoveries in the real data are unknown, we further tested Dynamic Filter on serials of synthetic datasets. The synthetic datasets were generated by varying four parameters systematically: lights and $i_{1/2}$ being smoothly or abruptly changed, abnormalities being continuously or discontinuously distributed, and error ratio being low or high. Furthermore, we added variations representing abnormalities and biological discoveries (dif-




(a) Phenotype of ELY and Col-0

(b) $i_{1/2}$ learned by our method

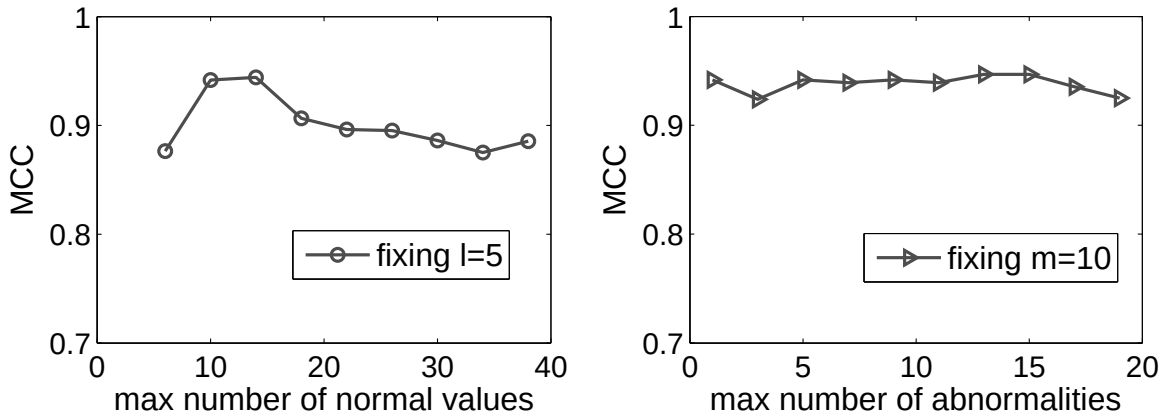
(c) ELY observed & modelled data in the coarse phase

(d) ELY observed & modelled in the refined phase

Figure 3.7: A case study on the real data shows that Dynamic Filter identifies true biological discoveries under the diurnal light condition. Lines with the same marker represent biological replicates.

ferent $i_{1/2}$ values) in the synthetic datasets. In total, 63 kinds of synthetic datasets in 9 groups were generated, and for each kind of synthetic data, we repeatedly generated 100 datasets.

Fig 3.9 shows the robustness of Dynamic Filter on different synthetic datasets generated under 9 different settings. The performance is evaluated using Matthews Correlation Coefficient on both abnormalities and on biological outliers. Each figure represents synthetic data generated under different settings (see details in supplementary section 2). Each point




(a) Performance vs. max number of normal values (b) Performance vs. max number of abnormalities

Figure 3.8: Performance test on temporal checking region size. (a) Fixing max number of abnormalities $n$, and varying $m$; (b) Fixing $m$, and varying max number of abnormalities $n$.

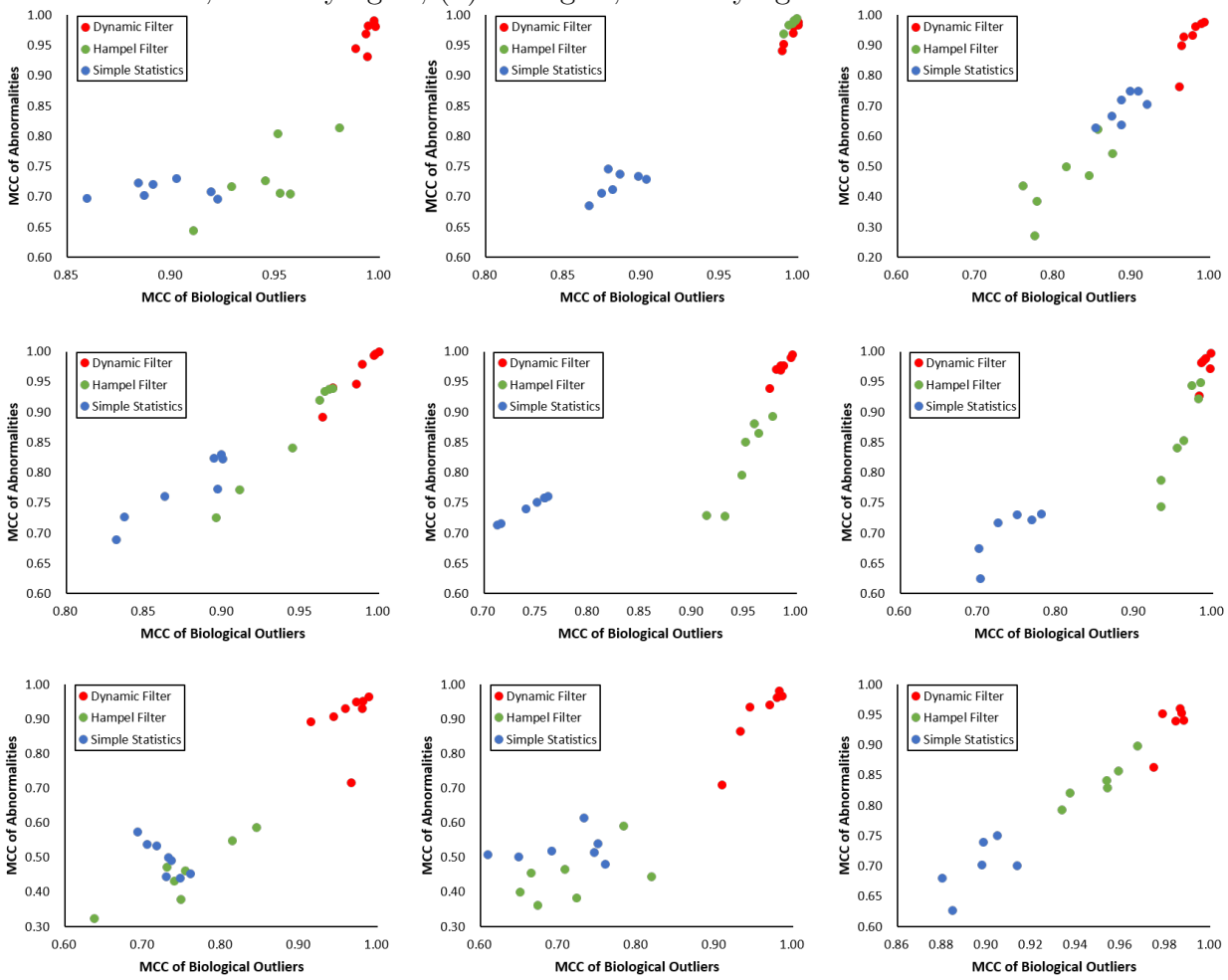



Figure 3.9: The Matthews Correlation Coefficient of biological discoveries and abnormalities on synthetic data.

in Fig 3.9 represents a MCC score of biological discovery identification at x-axis and a MCC score of abnormality identification at y-axis. The highest possible value is $(1.0, 1.0)$. The experimental results show that Dynamic Filter (red circle) is better than the other two methods in almost all the synthetic datasets. This is because Dynamic Filter can identify and remove abnormalities while reserving biological discoveries (see supplementary Table 1 and Table 2 for performance comparison on MCC and true positive rate respectively).

# Chapter 4

# Inter-functional analysis by PhenoCurve

In this Section, we propose a model based Bayes curve fitting algorithm, which is able to study both the values and the changing rates of the dynamic phenomics data. In particular, this algorithm is designed for studying the hidden parameter of phenomics, but actually it is not exclusive to phemomics data but also may be extended to be used on other models.

The rest of the chapter is arranged as follows. Section 4.1 motivates the problem and main intuition behind the proposed algorithm, the detailed description of the proposed algorithm. Section 4.2 presents the detailed implementation issues, the experimental setup, results and analysis.

## 4.1 Method

In order to explore the dynamic phenotype-environment relationships without the technical limitations in curve fitting and Bayesian NIG methods, we present a comprehensive data analysis approach *PhenoCurve* based on Bayesian theorem and polynomial generalization.

PhenoCurve has four components. First, it splits the whole phenotype and environment data into highly overlapped temporal regions using a sliding window approach, allowing for modeling the gradual change of $i_{1/2}$. Second, it employs a non-linear curve fitting method

to compute $i_{1/2}$ for each temporal region, and classifies the results into two groups, i.e. reliable ($D_r$) and unreliable ($D_u$), based on $R^2$. Third, using data from $D_r$, it estimates $i_{1/2}$ for each unreliable region in $D_u$ with polynomial generalization. Finally, it optimizes the $i_{1/2}$ values for all unreliable regions with Bayesian theorem using local data, resulting in increased reliability in curve fitting. We will introduce all the four steps in the following content. An illustrative example of $\Phi_{II}$ and light data is shown in Fig. 4.1AB.

### 4.1.1 Data Separation with Sliding Window

Due to biological constraints, the sampling rate of the phenotype is usually much lower than that of the environmental factors [14]. Subsequently, we split the whole dataset into highly overlapped temporal regions solely based on the phenotype data. In the fix-size sliding window approach, each temporal region has $d$ pairs of phenotype and environment values, and the values in each pair are probed at exactly the same time (or are close enough to each other). Each temporal region share $d-1$ values with the previous and the next region. The window width $d$ satisfies two conditions: 1) $i_{1/2}$ remains relatively constant within each temporal region; and 2) there are enough data in each temporal region for inferring $i_{1/2}$.

### 4.1.2 Local Curve Fitting

The next step is to infer $i_{1/2}$ for every temporal region $R_n$ centered around time point $t_n$. Its half light parameter $i^n_{1/2}$ can be estimated using least square curve fitting [20]. The general idea is to identify parameters in a fitting function to minimize the sum of all the square of errors between the predicted and the observed values. In our case, assuming there are $d$

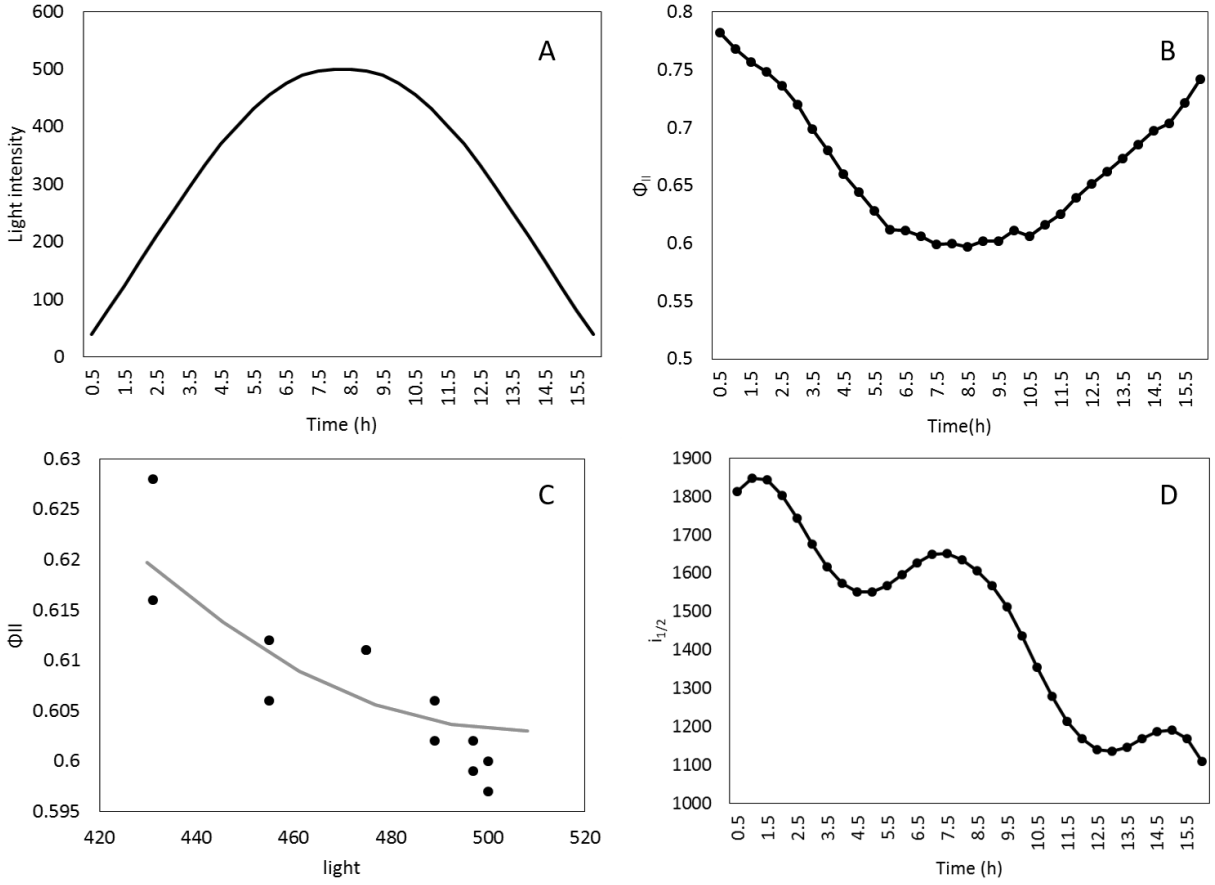



Figure 4.1: An illustrative example of PhenoCurve that optimizes both local fitting and smoothness of $i_{1/2}$, which shows A) dynamic light variation over time, B) corresponding $\Phi_{II}$ values, C) an optimized fitting on local region, and D) $i_{1/2}$ values on all temporal regions.

phenotype-environment pairs $\{(\Phi_{II}^1, i^1), (\Phi_{II}^2, i^2), \ldots, (\Phi_{II}^d, i^d)\}$ in $R_n$, we identify $i_{1/2}^n$ by:

$$i_{1/2}^n = \underset{i_{1/2}}{\operatorname{argmin}} \sum_{k=1}^{d} \left( \Phi_{II}^k - \frac{\max(\Phi_{II})}{1 + \frac{i^k}{i_{1/2}}} \right)^2 \tag{4.1}$$

The fitting procedure in Eq. 4.1 also yields a $R^2$ score, indicating the level of reliability of the fitting. Based on $R^2$, we divide all the temporal regions into two groups, i.e. reliable data $D_r$ and unreliable data $D_u$. In our experiment, $R^2$ threshold at 0.9 is used on both the real and the synthetic datasets.

### 4.1.3 Polynomial Generalization

Using the $i_{1/2}$ values for the reliable temporal regions in $D_r$ as inputs, we build a regularized polynomial linear regression model, aiming to generalize a $p$-order polynomial smooth curve of $i_{1/2}$ [63]. In a regularized polynomial linear regression model, order of one represents a linear model, order of two represents a quadratic form, and order of three and above works for arbitrary shapes. In our experiments, the highest order of polynomial term was set to three by using cross-validation. Note that using high-order polynomial would risk in over-fitting [7].

Let $\mathbf{X} = (1, t, i, \Phi_{II}, t^2, i^2, \Phi_{II}^2, t^3, i^3, \Phi_{II}^3)^T$ be the vector of polynomial terms, and $\mathbf{W} = (w_0, w_1, ..., w_9)^T$ be the coefficient vector of $\mathbf{X}$, the generalized polynomial linear regression model is $i_{1/2} = \mathbf{W}^T\mathbf{X}$.

The aim of the regularized fitting is to minimize $E(\mathbf{W})$ where

$$\begin{aligned} E(\mathbf{W}) &= \sum_{n=1}^{|D_r|}(\mathbf{W}^T\mathbf{X}_n - i_{1/2}^n)^2 + \lambda|\mathbf{W}|_2^2 \\ &= |\mathbf{W}^T\mathbf{X} - \mathbf{i_{1/2}}|_2^2 + \lambda|\mathbf{W}|_2^2 \end{aligned} \tag{4.2}$$

where $|D_r|$ is the number of reliable temporal regions identified in Section 4.1.2, $\mathbf{X}$ is the polynomial feature matrix, $\mathbf{i_{1/2}}$ is the vector of $i_{1/2}$ values learned from the reliable temporal regions, and $|\cdot|_2^2$ denotes $L_2$ norm of a vector.

In order to solve Eq. 4.2, we set $\frac{\partial E(\mathbf{W})}{\partial \mathbf{W}} = 0$, which yields

$$W_* = (\lambda\mathbf{I} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{i_{1/2}} \tag{4.3}$$

where $\lambda$ is a regularization term provided by user.

Applying $i_{1/2} = \mathbf{W}_*^T\mathbf{X}$ to data in every temporal region, including the unreliable regions, we therefore obtain a generalized predicted half-saturation parameter for every region $R_n$, denoted as $\hat{i}_{1/2}^n = W_*^T\mathbf{X}_n$. The generalized $i_{1/2}$ values ensure smooth hidden state variables under dynamic environments.

### 4.1.4 Bayesian MLE optimization

Given an unreliable temporal region $R_n$ in $D_u$, we optimize its local half light parameter $\hat{i}_{1/2}^{*n}$ using Bayesian theorem, such that $\hat{i}_{1/2}^{*n}$ fits best with both the local data in $R_n$ and $\hat{i}_{1/2}$ learned from the generalized model in the previous step. Mathematically, we look for optimal $\hat{i}_{1/2}^{*n}$ that satisfies

$$\hat{i}_{1/2}^{*n} = \underset{i_{1/2}}{\text{argmax}}\ P(i_{1/2}|R_n, \hat{i}_{1/2}) \tag{4.4}$$

Since $R_n$ and $\hat{i}_{1/2}$ are independent, we adopt the Bayesian theorem to solve Eq. 4.4:

$$\begin{aligned} P(i_{1/2}|R_n, \hat{i}_{1/2}) &= \frac{P(i_{1/2}, \hat{i}_{1/2}) \cdot P(R_n|i_{1/2}, \hat{i}_{1/2})}{P(R_n, \hat{i}_{1/2})} \\ &= \frac{P(i_{1/2}, \hat{i}_{1/2}) \cdot P(R_n|i_{1/2}, \hat{i}_{1/2})}{P(R_n)P(\hat{i}_{1/2})} \\ &= \frac{P(i_{1/2}|\hat{i}_{1/2}) \cdot P(R_n|i_{1/2})}{P(R_n)} \\ &= \frac{P(i_{1/2}|\hat{i}_{1/2}) \cdot \prod_{k=1}^{d} P(d_k|i_{1/2})}{P(R_n)} \\ &\propto P(i_{1/2}|\hat{i}_{1/2}) \cdot \prod_{k=1}^{d} P(d_k|i_{1/2}) \end{aligned} \tag{4.5}$$

where $d_k$ is the $k$th pair of phenotype-environment pair in $R_n$ and $R_n \in D_u$.

In Eq. 4.5, $P(i_{1/2}|\hat{i}_{1/2})$ represents the probability of $i_{1/2}$ given $\hat{i}_{1/2}$. We assuming $i_{1/2}$ follows Gaussian distribution with $i_{1/2} \sim \mathcal{N}(\hat{i}_{1/2}, \sigma_1^2)$, where $\hat{i}_{1/2}$ is the mean and $\sigma_1^2$ is the variance. In Eq. 4.5, $P(d_k|i_{1/2})$ represents the fitness of the phenotype-environment data given $i_{1/2}$. In fact, when $i_{1/2}$ is provided, $\Phi_{II}$ can be computed using Eq. 2.2. With the same Gaussian distribution assumption, we have $P(d_k|i_{1/2}) = \mathcal{N}(\Phi_{II}^k, \sigma_2^2)$. Here both $\sigma_1^2$ and $\sigma_2^2$ are user defined parameters.

By adopting the Gaussian distribution assumption on $P(i_{1/2}|\hat{i}_{1/2})$ and $P(d_k|i_{1/2})$, and by taking a log transform on Eq. 4.5, we have:

$$\begin{aligned} \ln Pr(i_{1/2}|R_n, \hat{i}_{1/2}) &\propto \ln \mathcal{N}(t|\hat{i}_{1/2}, \sigma_1^2) + \sum_{k=1}^{d} \ln \mathcal{N}(\Phi_{II}^k|f_{PI}(t_k, i_{1/2}), \sigma_2^2) \\ &\propto -\frac{(i_{1/2} - \hat{i}_{1/2})^2}{2\sigma_1^2} - \sum_{k=1}^{d} \frac{(\Phi_{II}^k - \frac{max(\Phi_{II})}{(1+i_k/i_{1/2})})^2}{2\sigma_2^2} \end{aligned} \tag{4.6}$$

Substituting the right part of Eq. 4.4 with Eq. 4.6, the final equation that can be solved by adopting the maximum likelihood estimation (MLE) is:

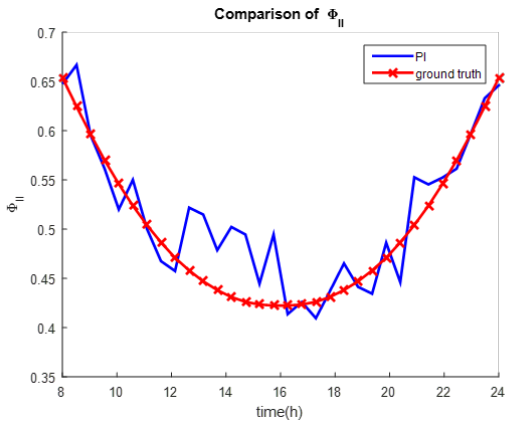
$$\hat{i}^{*n}_{1/2} = \underset{i_{1/2}}{\operatorname{argmin}} \left( \frac{(i_{1/2} - \hat{i}_{1/2})^2}{2\sigma_1^2} + \sum_{k=1}^{d} \frac{(\Phi^k_{II} - \frac{max(\Phi_{II})}{(1+i_k/i_{1/2})})^2}{2\sigma_2^2} \right) \tag{4.7}$$

Note that in order to increase the reliability of curve fitting in unreliable region $R_n$, during the Bayesian MLE optimization process, we search for the phenotype-environment pairs outside $R_n$ that have the best match to the fitting curve of $R_n$, and add them into $R_n$. Fig. 4.1CD shows an example of the fitting, which optimizes both local fitting and the smoothness of $i_{1/2}$.

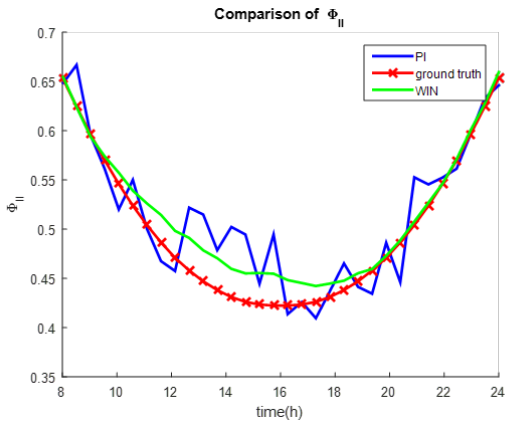
To demonstrate the PhenoCurve procedure, Fig. 4.2 shows the different stages of curve fitting. Fig. 4.2(a) presents the ground truth data for fitting, and as well as the observed input data which is composed of noise rate at 0.1 and bias at 0.1. The fitting result from sliding window approach(WIN) is shown in Fig. 4.2(b). Based on sliding window result, the data region is separated into reliable regions and unreliable regions as shown in Fig. 4.2(c). Lastly, PhenoCurve use the reliable regions to fit generalized half-saturation parameters(as shown in green curve with mark in Fig. 4.2(d)) and make the prediction on the unreliable regions(as shown in green curve without mark in Fig. 4.2(d)).

## 4.2 Experimental Results

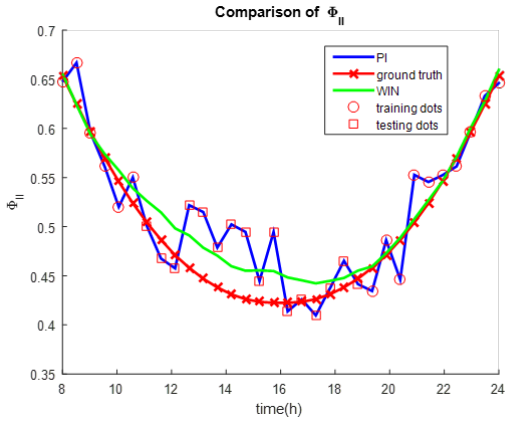
We evaluated the performance of PhenoCurve on both the real and simulated phenotype data in terms of fitting performance and fitting reliability. We also compared PhenoCurve with five existing methods i.e. 1) the direct computation with PI function (PI), 2) the one-



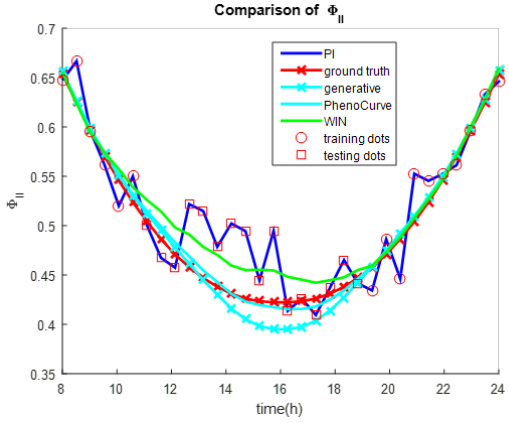

(a) Ground Truth and Input data.




(b) Sliding window approach fitting (WIN).




(c) Separation of training and testing regions.




(d) PhenoCurve gneralization and Bayes optimization fitting.

Figure 4.2: Demonstration of PhenoCurve running example on synthetic data.

window curve fitting (ONE), 3) the sliding-window based curve fitting (WIN), 4) the kernel smoothing method using local linear regression (LLR), and 5) the Bayesian linear model with normal inverse gamma prior (NIG). All these methods have been introduced in the Related Work Section.

### 4.2.1 Experimental Data

We first tested the performance of PhenoCurve using the real plant photosynthetic phenotype data consisting of 375 Arabidopsis thaliana plants (330 confirmed T-DNA insertion mutants

and 45 wild type plants) [3, 2]. During the experiment, all the plants were evenly sampled at 32 time points. All the environmental factors except light are constant. Following a sinusoidal curve, light intensity changes gradually from $35\mu molm^{-2}s^{-1}$ to $500\mu molm^{-2}s^{-1}$ then goes back to $35\mu molm^{-2}s^{-1}$ during the experiment. The photosynthetic phenotype values vary dramatically across plants, reflecting potential differences in development, stress responses or regulation of processes such as stomatal conductance, photodamage and storage of photosynthate [32, 14].

The synthetic data were generated in three steps. First, we randomly defined a vector of $i_{1/2}$ that changes gradually over time. Second, we reconstructed a vector of phenotype $\Phi_{II}$ using the vector of $i_{1/2}$, the same vector of light as the real data, and the PI function [38]. Third, we randomly added noise and bias (levels vary from 5% to 15%) to the phenotype data. The process was repeated 4,000 times to generate the full synthetic data (see Table S1).

### 4.2.2 Evaluation Criteria

We define four performance evaluation criteria, and applied all of them on both the unreliable regions $D_u$ and the reliable regions $D_r$, as well as the whole regions of the phenotype data $D$.

First, coefficient of determination, denoted as $R^2$, is often used as the main criteria for measuring whether a curve fitting is adequate [28, 10]. In our case, $R^2$ is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\Phi_{II}(t_i) - \hat{\Phi}_{II}(t_i))^2}{\sum_{i=1}^{n}(\Phi_{II}(t_i) - \overline{\Phi_{II}})^2} \tag{4.8}$$

where $\hat{\Phi_{II}}(t_i)$ is the fitted $\Phi_{II}$ value at time $t_i$, and $\overline{\Phi_{II}}$ is the averaged $\Phi_{II}$ values in a

temporal region. In Eq. 4.8, $R^2$ measures the fraction of the total variation in the phenotype data that can be explained by the curve. Higher values indicate that the curve fits the data better. If $R^2 = 1.0$, all points lie exactly on the curve with no scatter.

Second, we compute the smoothness of the $i_{1/2}$ vector. For a continuous curve, smoothness can be measured using high order of derivatives. For discrete values (which is our case), we measure all the angles formed by adjacent temporal regions. Mathematically, we have:

$$smoothness(i_{1/2}) = \frac{1}{|D|} \sum_{n=1}^{|D|} [\alpha_n \leq T_\alpha] \tag{4.9}$$

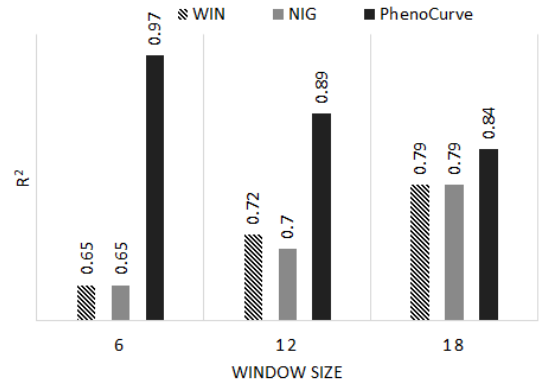
where $D$ is a set of temporal regions, $\alpha_n = |\arctan(i_{1/2}^{n+1} - i_{1/2}^{n}) - \arctan(i_{1/2}^{n} - i_{1/2}^{n-1})|$ represents the angle difference centered around temporal region $R_n$ ($R_n \in D$), $T_\alpha$ is a user given angle threshold, and $[X]$ returns 1 if the condition $X$ is satisfied, otherwise it returns 0. In our experiment, $T_\alpha = 30^o$.

Finally, for synthetic data, we computed both $\Delta\Phi_{II}$ and $\Delta i_{1/2}$, which are the sum of all the absolute differences between every phenotype value and its corresponding value on the fitted curve and the sum of all the absolute differences between every $i_{1/2}$ value and its corresponding parameter of the fitted PI curve respectively.

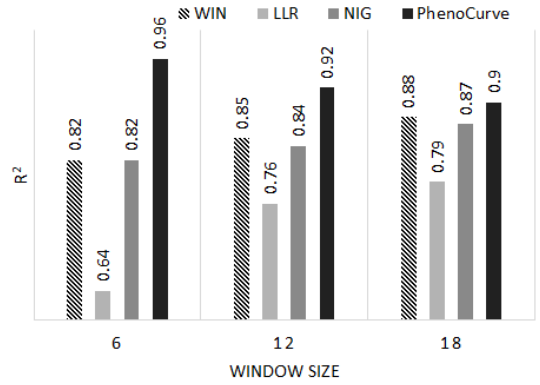
### 4.2.3 Experimental Results on Real Data

We ran PhenoCurve on the real photosynthesis phenotype data using three different window sizes, i.e. 6, 12 and 18. We compared the performance of PhenoCurve with five existing methods using the same common parameters.

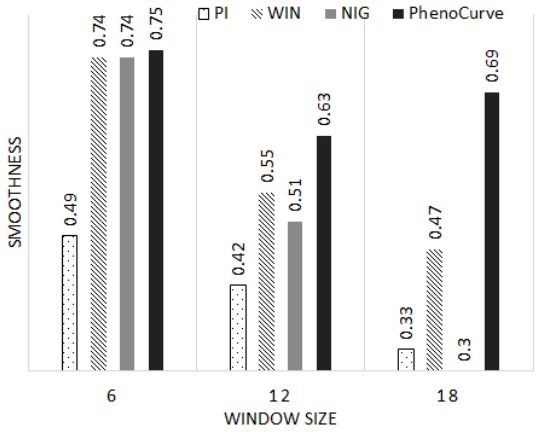
The experimental results indicate that PhenoCurve has significantly more reliable curve fitting results. Fig. 4.3 shows that on the unreliable regions, the coefficient of determination



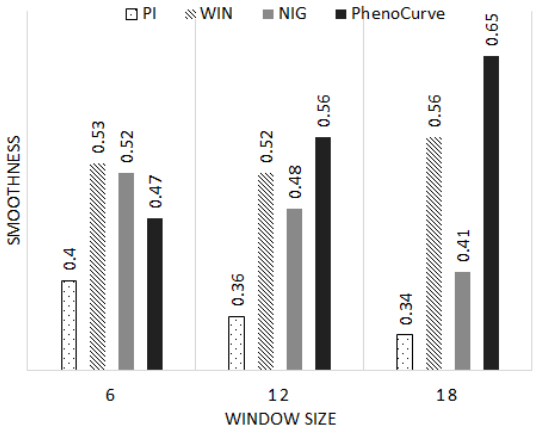

(a) $R^2$ on unreliable regions.




(b) $R^2$ on whole data.



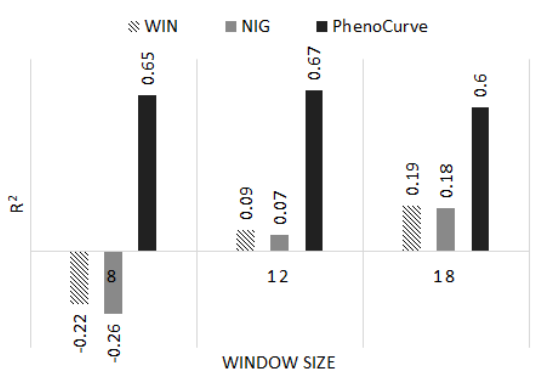

(c) Smoothness on unreliable regions.



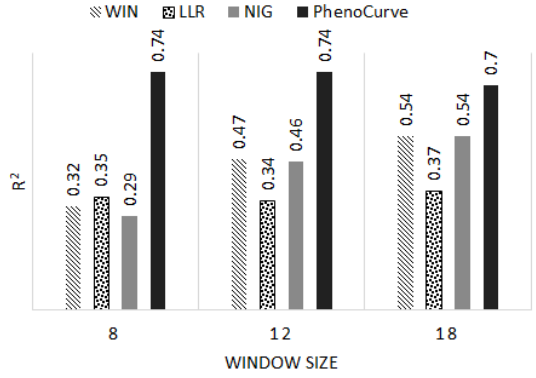

(d) Smoothness on whole data.

Figure 4.3: Coefficient of determination $R^2$ and smoothness of $i_{1/2}$ on the unreliable regions and the whole real phenotype data.

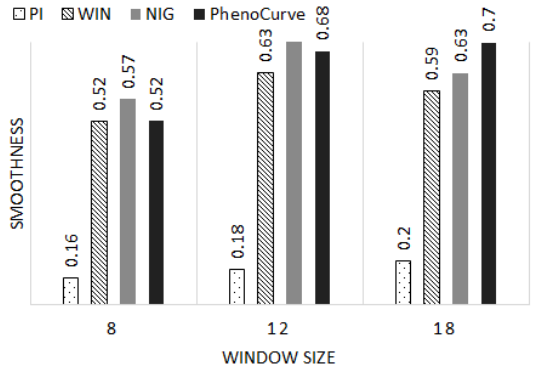
($R^2$) increases 23% from 0.72 to 0.89 when window size is 12, compared with the sliding-window curve fitting method. Meanwhile, the smoothness of the fitted curve increases from 0.55 in sliding-window method to 0.63 in PhenoCurve approach. Among all tests on real data, PhenoCurve yields highest $R^2$ in both unreliable data and whole data parts. PhenoCurve also outputs smoothest curve in all cases except the one with window size = 6. Overall, PhenoCurve performs best in experiments on real data.



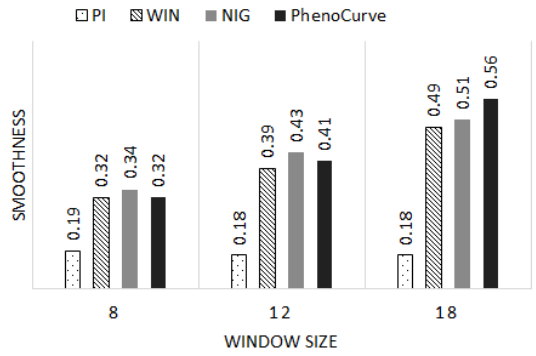

(a) $R^2$ on unreliable regions.




(b) $R^2$ on whole data.




(c) Smoothness on unreliable regions.



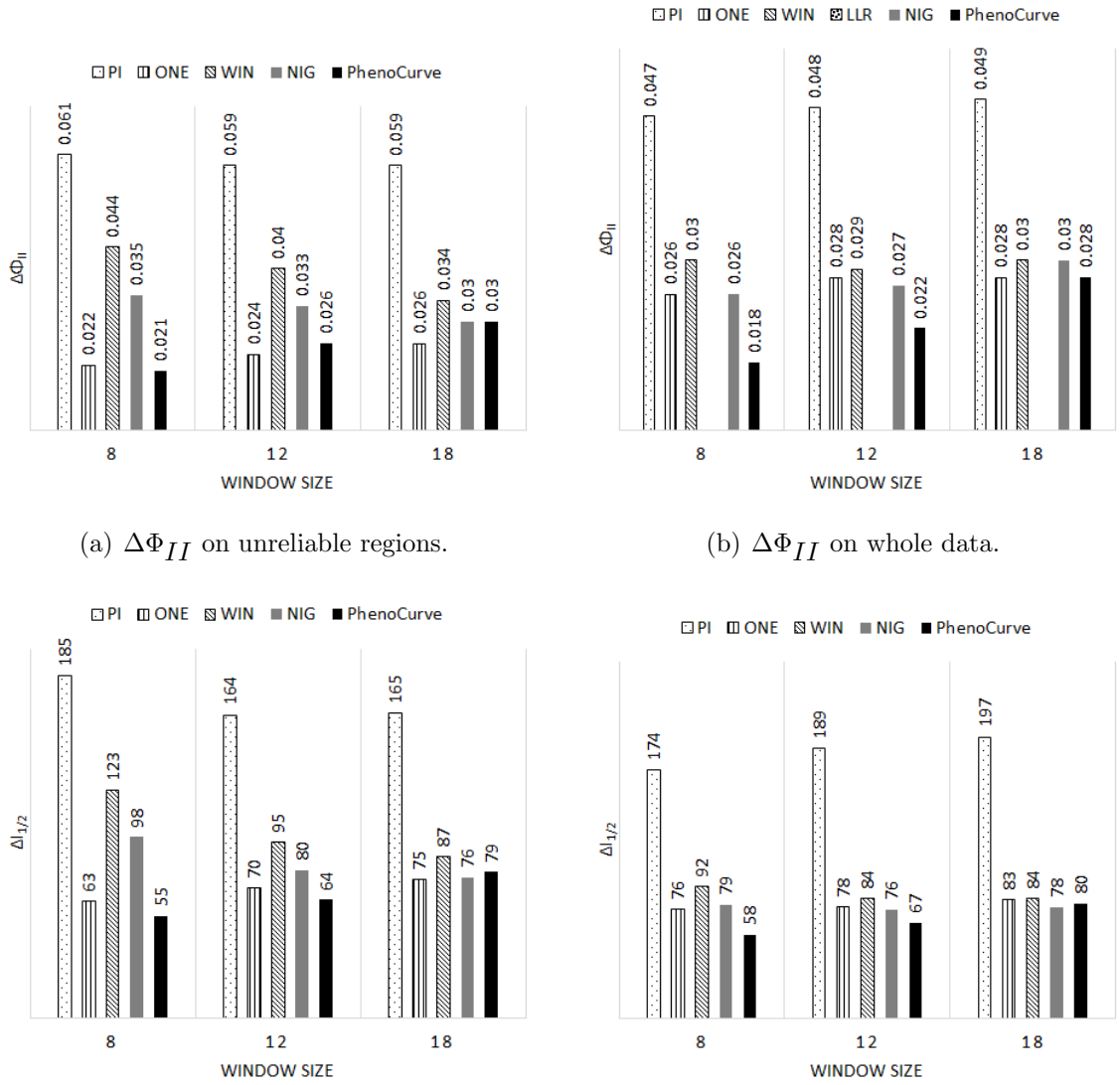

(d) Smoothness on whole data.

Figure 4.4: Coefficient of determination $R^2$ and smoothness of $i_{1/2}$ on the unreliable regions and the whole synthetic phenotype data.

### 4.2.4 Experimental Results on Synthetic Data

Similarly, we compared the performance of PhenoCurve and the five existing methods on a synthetic phenotype data using three different window sizes.

The experimental results indicate that PhenoCurve has most reliable and smoothest curve fitting results, and it has lowest error rate for $\Phi_{II}$ and $HL$ . Fig. 4.4 shows that on the unreliable regions, the coefficient of determination ($R^2$) increases significantly from barely 0.09 in sliding-window method to 0.67 in PhenoCurve when window size is 12. Besides, smoothness of PhenoCurve is always the best among all methods. Fig. 4.5 shows the comparison of




(a) $\Delta\Phi_{II}$ on unreliable regions.

(b) $\Delta\Phi_{II}$ on whole data.

(c) $\Delta i_{1/2}$ on unreliable regions.

(d) $\Delta i_{1/2}$ on whole data.

Figure 4.5: $\Delta\Phi_{II}$ and $\Delta i_{1/2}$ on the synthetic phenotype data with 0.10 noise and bias rate.

error rate $\Delta\Phi_{II}$ and $\Delta i_{1/2}$. On unreliable data regions, PhenoCurve performs better than all other methods in term of $\Delta i_{1/2}$. As for $\Delta\Phi_{II}$, PhenoCurve and one window approach has best performance in unreliable regions, while PhenoCurve performs best in whole data. This is due to the fact that PhenoCurve not only optimize unreliable data, but also optimize the reliable regions. Overall, PhenoCurve performs best on synthetic data.

### 4.2.5 Biological Verification

Given all the $i_{1/2}$ values of all the Arabidopsis mutant lines, we computed the maximal relative $i_{1/2}$ as $max|i_{1/2}(t_m) - i_{1/2}(t_n)|$ for all time points $t_m$ and $t_n$ satisfying $i(t_m) = i(t_n)$ and $m \neq n$. The purpose is to check the recovery ability of a plant by probing it before and after light stress. The distribution of data in Fig. 4.6 shows that we can easily cluster all the Arabidopsis mutant lines into three groups $A$, $B$, $C$ as indicated in the figure. Mutants in group A and B have high photosynthesis, whereas group C mutants have relatively low photosynthesis. Mutants in group A and C are more resistant to dynamic light stress than group B mutants.

In order to verify the biological discovery, a biological experiment by measuring $qI$ (photoinhibition) of the same plants using DEPI [14] may potentially be designed. In summary, learning $i_{1/2}$ with PhenoCurve can greatly simplify the process to identify the complicated phenotype-environment relationships and to visualize the results, enabling biologists to discover the mechanism that regulate plants in response to dynamic environment.

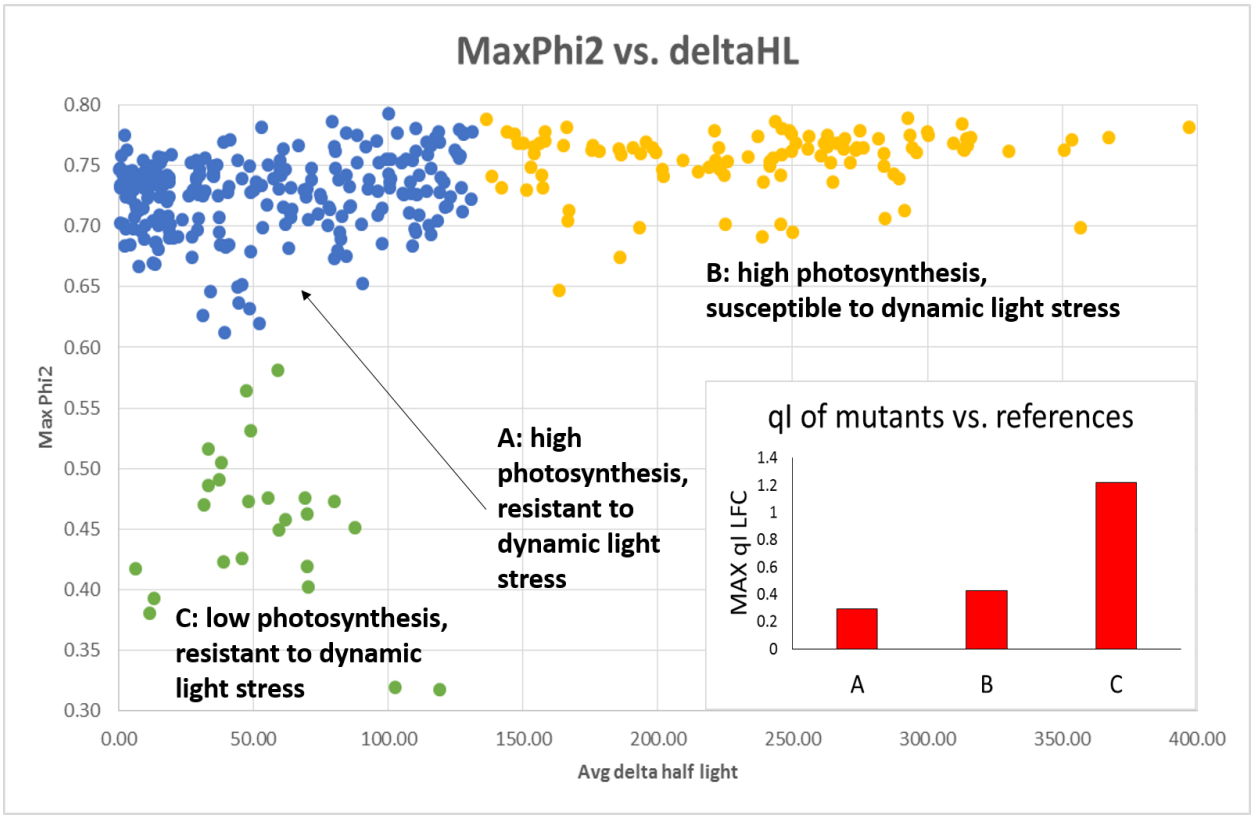



Figure 4.6: Phenotype clustering based on maximal relative $i_{1/2}$ and maximal $\Phi_{II}$.

# Chapter 5

# Summary and Conclusions

In this thesis, we designed two techniques for control data quality and explore data of phenotyping values. The proposed applications, namely, Dynamic Filter and PhenoCurve, achieve the ultimate task around two questions including

- How can we get high quality data from problematic raw data?
- How can we do curve fitting for dynamically changing phenotyping values with high reliability?

## 5.1 Contributions

This thesis mainly answers the two questions raised above by proposing specific applications as follows, giving theoretical studies and providing real case and synthetic case tests.

### 5.1.1 Data cleaning for high data quality

With an aim towards identifying targets for improving energy yield, advanced technologies in high-throughput plant photosynthetic phenotyping have been developed ([29, 14]). These systems can be used to quantify photosynthetic behavior in genetically diverse populations and to draw relationships among genotype, phenotype and biological function, leading to better understanding of the underlying mechanisms that control the photosynthetic proper-

ties under various environmental conditions ([52, 19]). As a consequence of the long-time high-throughput plant phenotyping, the scale of plant phenomics data grows exponentially. However, the quality of phenotype data may be skewed by sources of noise that are difficult to remove in the data collection step.

The purpose of plant phenotyping is to discover phenotype values that are significantly different from a reference. But phenotype values leading to biological discoveries maybe obscured by abnormal values caused by errors during detection. To ensure high data quality, effective data cleaning should be considered a primary task. However, since advanced data cleaning algorithms are primarily based on indiscriminate outlier detection, they may remove both abnormalities and biological discoveries not separable in the data distribution.

We have developed a new coarse-to-refined model called *Dynamic Filter* to effectively identify both abnormalities and biological discoveries by adopting a widely-used photosynthetic model. Specifically, Dynamic Filter is a residual analysis approach by dynamically tracing statistical distributions of all samples rather than individuals, and incorporating EM for performance optimization in refined checking regions.

We note that, certain events, such as transient changes in growth environment, could introduce signals similar to growth lighting malfunction, which could be wrongly labeled as abnormalities by Dynamic Filter. Therefore, instead of automatically deleting all the predicted abnormalities, we send all of them to domain experts for confirmation. Meanwhile, all raw data are kept for any rollback operation.

Experimental results show that our model is significantly better than the existing data cleaning tools on both real phenomics data and synthetic data. Dynamic Filter may have a wide impact because of the rapid increase of large-scale phenotyping technologies. It should be noted that although we used a photosynthesis-specific curve, the model itself is

independent of actual biological constraints. In principle, our approach can be used to clean data for any number of phenotypes as long as suitable theoretical curves can be derived for their behavior. Implementation for new use cases would involve substituting the appropriate theoretical curve into the program, calculating the residuals of fits to the data sets, and optimizing the fitting procedure (as in Fig 3.1).

### 5.1.2 Inter-functional analysis

Photosynthesis phenotypes are usually measured under dynamic conditions, in a relatively long time period, and on plants with vastly different genetic backgrounds. To meet the growing needs to study the dynamic relationships between phenotypes and environmental factors using limited biological knowledge, we developed a new tool called PhenoCurve.

PhenoCurve splits the whole phenotype and environment data into highly overlapped temporal regions, employs non-linear curve fitting methods to calculate $i_{1/2}$ for the reliable regions, and then optimize the $i_{1/2}$ for the unreliable regions using polynomial generalization and Bayesian MLE. The results on both synthetic and real data show that PhenoCurve os significantly better than the other five existing methods.

Table 5.1: Testing the robustness of PhenoCurve against fixed window approach with multiple noise and bias rates.

| | **Unreliable data** | | | | **Whole data** | | | |
|---|---|---|---|---|---|---|---|---|
| noise and bias | $R^2 \uparrow$ | smoothness $\uparrow$ | $\Delta\Phi_{II} \downarrow$ | $\Delta HL \downarrow$ | $R^2 \uparrow$ | smoothness $\uparrow$ | $\Delta\Phi_{II} \downarrow$ | $\Delta HL \downarrow$ |
| 0.05 | 67% | 7% | 83% | 35% | 24% | 20% | 33% | 15% |
| 0.10 | 87% | 7% | 54% | 50% | 36% | 5% | 32% | 25% |
| 0.15 | 118% | 3% | 46% | 42% | 46% | 5% | 27% | 22% |

In order to test the robustness of PhenoCurve, we ran PhenoCurve on synthetic data with different levels of noise and bias. The results shown in Table. 5.1 indicates that PhenoCurve is more effective than fix window approach with the increase of noise and bias rate. Compared to fix window approach, PhenoCurve yields hight $R^2$, smoother curve, lower error in $\Phi_{II}$

and half light parameters.

## 5.2 Conclusions

This dissertation answers the two questions raised in the beginning of this chapter by presenting machine learning techniques in data quality control and data exploration of dynamically changing phenotyping values. To address these problems, Dynamic Filter and PhenoCurve are proposed to clean raw data and explore cleaned data.

The concluded research makes significant contributions to (i) the realistic solution for maintain high quality and clean problematic data from raw input, (ii) the challenges of exploring and obtaining reliable and comparable hidden parameters from phenotyping data.

# Chapter 6

# Future Work

## 6.1 Future Work

The studies presented in the thesis lead to several important research questions, which may be extended in the future.

### 6.1.1 Dynamic Filter for data quality control

Future work may incorporate semi-supervised learning techniques into the application. Current version of Dynamic Filter process result with one pass, and does not interact with human verification to automatically adjust the parameter setting.

As for application potential, as long as a suitable theoretical curve is given, the theoretical photosynthetic curve can be updated with the new one, and residuals of the curve can be learned at the coarse level, and then be optimized in the refined level (see Fig 3.1). Therefore, it is possible to automate the cleansing process on any long time-series data for a variety of applications.

### 6.1.2 Inter-functional analysis

In the future, we will extend PhenoCurve to model multiple phenotypes and multiple environmental factors. One of the technical challenges is that with the increase of data types,

the number of parameters will rapidly increased, which either requires significantly more sampling data, or requires a better algorithm to model the data.

# APPENDIX

# Pseudo-code of Dynamic Filter

**Algorithm 4** DynamicFilter: Dynamic Filter Algorithm

**Input:**

- $\{\Phi_{II}\}$: array of photosynthesis intensity , Light,99%
- *light*: array of light intensity
- 99%: 99% is the user-specified confidence interval

$[RSD, I_{i/2t}] \leftarrow$ PI_CurveFitting($\{\Phi_{II}\}, Light$) 2.1
$[\mu_{main}, \sigma_{main}] \leftarrow GMM(RSD)$
$[r_{min}, r_{max}] \leftarrow$ get Confidence Interval($\mu_{main}, \sigma_{main}, 99\%$) 5
$C \leftarrow$ Update Candidate($RSD, r_{min}, r_{max}$) 6
**repeat**
    $\Psi \leftarrow (\{\Phi_{II}\}, \{RSD\}, Light)$
    $C \leftarrow$ KNN($\Psi, C, k_0$) 1
    $\{\Phi_r\} \leftarrow$ Update Window ($C, num$) 7
    **for** $\Phi_r$ in $\{\Phi_r\}$ **do**
        $\Phi_r \leftarrow$ EM_Optimization($\Phi_r, Light, 99\%$) 3
    **end for**
    $C \leftarrow$ Consensus(W) 8
**until** $C$ is stable
**Output:** $C$

**Algorithm 5** Sub-procedures: get Confidence Interval

**Input:**

- $\mu$: mean of Gaussian distribution
- $\sigma$: standard deviation of Gaussian distribution
- 99%: user chosen confidence interval, corresponds to $3\sigma$

$r_{min} \leftarrow \mu - 3 \cdot \sigma$
$r_{max} \leftarrow \mu + 3 \cdot \sigma$
**Output:** $[r_{min}, r_{max}]$

**Algorithm 6** Sub-procedures: Update Candidate

**Input:**

- $R$: array of residual
- $r_{min}$: minimum residual threshold
- $r_{max}$: maximum residual threshold

**for** $r$ in $R$ **do**
  **if** $r_{min} \leq r \leq r_{max}$ **then**
    $C_r \leftarrow 0$
  **else**
    $C_r \leftarrow 1$
  **end if**
**end for**
**Output:** $C$

**Algorithm 7** Sub-procedures: Update Window

**Input:**

- $C$: candidate array
- $num$: maximum number of abnormality in each region

**for** abnormality $c$ in $C$ **do**
  **repeat**
    $w.abnormality(end + 1) \leftarrow c$
    next $c$
  **until** $length(w.abnormality) >= num$
  $w.normal \leftarrow$ [$num$ points ahead,$num$ points after]
  $W_i \leftarrow w$
  clear $w$
**end for**
**Output:** $W$

**Algorithm 8** Sub-procedures: Consensus

**Input:**

- $W$: array of windows

$(Count_n, Count_{ab}) \leftarrow ([], [])$
**for** $w$ in $W$ **do**
  **for** $Idx_n$ in $w.normal$ **do**
    $Count_n[Idx_n] \leftarrow Count_n[Idx_n] + 1$
  **end for**
  **for** $Idx_{ab}$ in $w.normal$ **do**
    $Count_{ab}[Idx_{ab}] \leftarrow Count_{ab}[Idx_{ab}] + 1$
  **end for**
**end for**
$C[i] \leftarrow (Count_n[i] < Count_{ab}[i])$ for each $i$
**Output:** $C$

# Proof of Equation 2

Let $P = \frac{P_{max} \times [I]}{i_{1/2} + [I]}$, $[I] = i(t)$ and $P = \Phi_{II}(t) \times i(t)$, we have $\Phi_{II}(t) \times i(t) = \frac{P_{max} \times i(t)}{i_{1/2} + i(t)}$, so that:

$$\Phi_{II}(t) = \frac{P_{max}}{i_{1/2} + i(t)}$$

Given the definition of $P_{max}$ (the maximum potential photosynthetic rate per individual) and $i_{1/2}$ (the light intensity at which the photosynthetic rate proceeds at half $P_{max}$), we have:

$$\frac{P_{max}}{2} = \frac{\max(\Phi_{II})}{2} \times i_{1/2}$$

By combining both equations, we have:

$$\Phi_{II}(t) = \frac{\max(\Phi_{II})}{1 + \frac{i(t)}{i_{1/2}}}$$

Done.

# Performance comparison of MCC and true positive rate on synthetic datasets

This section presents the performance comparison of Dynamic Filter, Simple Statistics and Hampel Filter on different synthetic datasets generated under different settings. The figure results are shown in Figure 10 of the paper, and the numerical values are in Table 1 and Table 2.

The performance compared are Matthews Correlation Coefficient and True Positive Rate on abnormalities and on biological outliers. Here the outliers are values generated with half-light factors significantly different from the averaged one.

There are 9 groups of different settings, with each group containing 7 different cases varying the deviation degree of abnormalities. Details of the 9 groups of settings are:

1. Light intensity generated from 3 frequency FFT, half light factor changes smoothly along time, abnormalities points aggregate, abnormalities rate is 7.8%.

2. Light intensity generated from 3 frequency FFT, half light factor changes smoothly along time, abnormalities points occurs every other point, abnormalities rate is 7.8%.

3. Light intensity generated from 3 frequency FFT, half light factor fluctuates along time, abnormalities points aggregate, abnormalities rate is 7.8%.

4. Light intensity generated from 1 frequency FFT, half light factor changes smoothly along time, abnormalities points aggregate, abnormalities rate is 7.8%.

5. Light intensity generated from 3 frequency FFT, half light factor changes smoothly along time, abnormalities points aggregate, abnormalities rate is 14.1%.

6. Light intensity generated from 1 frequency FFT, half light factor changes smoothly along time, abnormalities points aggregate, abnormalities rate is 14.1%.

7. Light intensity generated from 3 frequency FFT, half light factor fluctuates along time, abnormalities points aggregate, abnormalities rate is 14.1%.

8. Light intensity generated from 3 frequency FFT, half light factor fluctuates along time, abnormalities points aggregate, abnormalities rate is 14.1%.

9. Light intensity generated from 3 frequency FFT, half light factor fluctuates along time, abnormalities points occur every other point, abnormalities rate is 7.8%.

Table 1: Performance comparison on 63 types of synthetic datasets. Each pair of the score is a MCC score and its standard deviation. Bold font indicates the best performance in each dataset.

| **Dataset** | **Simple** | | **Statistics** | | **Hampel** | | **Filter** | | **Dynamic** | | **Filter** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **outlier** | **Std** | **abnormality** | **Std** | **outlier** | **Std** | **abnormality** | **Std** | **outlier** | **Std** | **abnormality** | **Std** |
| 1 | 0.86 | 0.03 | 0.70 | 0.05 | 0.91 | 0.06 | 0.64 | 0.12 | **0.99** | 0.00 | **0.93** | 0.09 |
| 2 | 0.89 | 0.00 | 0.70 | 0.02 | 0.95 | 0.02 | 0.71 | 0.12 | **0.99** | 0.01 | **0.95** | 0.04 |
| 3 | 0.89 | 0.01 | 0.72 | 0.04 | 0.95 | 0.03 | 0.80 | 0.05 | **1.00** | 0.01 | **0.99** | 0.02 |
| 4 | 0.92 | 0.01 | 0.71 | 0.03 | 0.93 | 0.04 | 0.72 | 0.09 | **1.00** | 0.00 | **0.98** | 0.01 |
| 5 | 0.92 | 0.01 | 0.70 | 0.02 | 0.96 | 0.03 | 0.70 | 0.08 | **1.00** | 0.00 | **0.99** | 0.01 |
| 6 | 0.90 | 0.01 | 0.73 | 0.05 | 0.98 | 0.02 | 0.81 | 0.08 | **0.99** | 0.01 | **0.97** | 0.02 |
| 7 | 0.88 | 0.01 | 0.72 | 0.03 | 0.94 | 0.03 | 0.73 | 0.07 | **0.99** | 0.01 | **0.98** | 0.02 |
| 8 | 0.87 | 0.03 | 0.69 | 0.03 | **1.00** | 0.00 | **0.99** | 0.00 | 0.99 | 0.01 | 0.95 | 0.04 |
| 9 | 0.87 | 0.03 | 0.71 | 0.03 | **1.00** | 0.00 | **0.99** | 0.01 | 0.99 | 0.01 | 0.94 | 0.05 |
| 10 | 0.89 | 0.01 | 0.74 | 0.02 | 0.99 | 0.01 | 0.97 | 0.02 | **1.00** | 0.00 | **0.98** | 0.01 |
| 11 | 0.90 | 0.01 | 0.73 | 0.04 | **1.00** | 0.00 | **0.99** | 0.01 | 1.00 | 0.00 | 0.98 | 0.01 |
| 12 | 0.90 | 0.02 | 0.73 | 0.06 | **0.99** | 0.01 | **0.99** | 0.01 | 1.00 | 0.01 | 0.97 | 0.02 |
| 13 | 0.88 | 0.02 | 0.75 | 0.03 | **1.00** | 0.00 | **0.99** | 0.01 | 1.00 | 0.00 | 0.98 | 0.01 |
| 14 | 0.88 | 0.02 | 0.71 | 0.05 | **1.00** | 0.00 | **1.00** | 0.01 | 1.00 | 0.00 | 0.99 | 0.02 |
| 15 | 0.85 | 0.02 | 0.63 | 0.03 | 0.78 | 0.09 | 0.27 | 0.08 | **0.96** | 0.02 | **0.76** | 0.09 |
| 16 | 0.87 | 0.01 | 0.67 | 0.05 | 0.76 | 0.10 | 0.44 | 0.11 | **0.96** | 0.04 | **0.90** | 0.07 |
| 17 | 0.89 | 0.05 | 0.64 | 0.04 | 0.78 | 0.06 | 0.38 | 0.13 | **0.98** | 0.02 | **0.93** | 0.02 |
| 18 | 0.90 | 0.01 | 0.75 | 0.05 | 0.86 | 0.07 | 0.62 | 0.11 | **0.99** | 0.01 | **0.98** | 0.02 |
| 19 | 0.91 | 0.02 | 0.75 | 0.05 | 0.88 | 0.07 | 0.54 | 0.18 | **0.98** | 0.02 | **0.97** | 0.04 |
| 20 | 0.92 | 0.03 | 0.70 | 0.03 | 0.82 | 0.07 | 0.50 | 0.11 | **0.97** | 0.02 | **0.93** | 0.04 |
| 21 | 0.89 | 0.03 | 0.72 | 0.02 | 0.85 | 0.03 | 0.47 | 0.05 | **0.99** | 0.02 | **0.97** | 0.03 |
| 22 | 0.83 | 0.03 | 0.69 | 0.02 | 0.90 | 0.02 | 0.73 | 0.06 | **0.96** | 0.04 | **0.89** | 0.08 |
| 23 | 0.84 | 0.04 | 0.73 | 0.03 | 0.91 | 0.02 | 0.77 | 0.05 | **0.99** | 0.01 | **0.98** | 0.02 |
| 24 | 0.86 | 0.01 | 0.76 | 0.01 | 0.94 | 0.02 | 0.84 | 0.07 | **1.00** | 0.00 | **1.00** | 0.00 |
| 25 | 0.90 | 0.01 | 0.77 | 0.02 | 0.96 | 0.01 | 0.92 | 0.02 | **1.00** | 0.00 | **0.99** | 0.01 |
| 26 | 0.90 | 0.01 | 0.82 | 0.02 | 0.97 | 0.01 | 0.94 | 0.02 | **0.99** | 0.01 | **0.95** | 0.04 |
| 27 | 0.89 | 0.01 | 0.82 | 0.02 | 0.97 | 0.02 | 0.93 | 0.03 | **1.00** | 0.00 | **1.00** | 0.01 |
| 28 | 0.90 | 0.02 | 0.83 | 0.02 | 0.97 | 0.02 | 0.94 | 0.03 | **0.97** | 0.04 | **0.94** | 0.08 |
| 29 | 0.71 | 0.01 | 0.56 | 0.03 | 0.93 | 0.03 | 0.73 | 0.05 | **0.97** | 0.02 | **0.94** | 0.05 |
| 30 | 0.72 | 0.01 | 0.58 | 0.03 | 0.91 | 0.03 | 0.73 | 0.06 | **0.99** | 0.01 | **0.98** | 0.01 |
| 31 | 0.76 | 0.03 | 0.55 | 0.06 | 0.95 | 0.02 | 0.80 | 0.06 | **0.98** | 0.01 | **0.98** | 0.02 |
| 32 | 0.75 | 0.04 | 0.58 | 0.05 | 0.96 | 0.01 | 0.87 | 0.07 | **1.00** | 0.00 | **0.99** | 0.00 |
| 33 | 0.76 | 0.03 | 0.56 | 0.07 | 0.98 | 0.01 | 0.89 | 0.04 | **0.99** | 0.01 | **0.99** | 0.01 |
| 34 | 0.74 | 0.03 | 0.58 | 0.06 | 0.96 | 0.01 | 0.88 | 0.04 | **0.99** | 0.02 | **0.97** | 0.03 |
| 35 | 0.76 | 0.06 | 0.53 | 0.05 | 0.95 | 0.02 | 0.85 | 0.02 | **0.98** | 0.01 | **0.97** | 0.02 |
| 36 | 0.67 | 0.01 | 0.58 | 0.03 | 0.93 | 0.01 | 0.74 | 0.02 | **0.98** | 0.02 | **0.93** | 0.09 |
| 37 | 0.70 | 0.02 | 0.62 | 0.02 | 0.93 | 0.01 | 0.79 | 0.04 | **0.99** | 0.02 | **0.99** | 0.02 |
| 38 | 0.70 | 0.03 | 0.67 | 0.02 | 0.95 | 0.03 | 0.84 | 0.03 | **0.99** | 0.02 | **0.98** | 0.02 |
| 39 | 0.72 | 0.03 | 0.72 | 0.02 | 0.96 | 0.01 | 0.85 | 0.04 | **0.99** | 0.02 | **0.98** | 0.02 |
| 40 | 0.75 | 0.03 | 0.73 | 0.01 | 0.97 | 0.01 | 0.94 | 0.02 | **1.00** | 0.00 | **1.00** | 0.01 |
| 41 | 0.77 | 0.03 | 0.72 | 0.02 | 0.98 | 0.01 | 0.95 | 0.03 | **0.99** | 0.02 | **0.99** | 0.02 |
| 42 | 0.78 | 0.02 | 0.73 | 0.03 | 0.98 | 0.01 | 0.92 | 0.06 | **1.00** | 0.00 | **0.97** | 0.05 |

Table. 1 (cont'd)

| Dataset | Simple | | Statistics | | Hampel | | Filter | | Dynamic | | Filter | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **outlier** | **Std** | **abnormality** | **Std** | **outlier** | **Std** | **abnormality** | **Std** | **outlier** | **Std** | **abnormality** | **Std** |
| 43 | 0.73 | 0.03 | 0.44 | 0.07 | 0.64 | 0.08 | 0.32 | 0.04 | **0.91** | 0.02 | **0.89** | 0.01 |
| 44 | 0.74 | 0.09 | 0.49 | 0.06 | 0.75 | 0.05 | 0.38 | 0.03 | **0.94** | 0.03 | **0.91** | 0.04 |
| 45 | 0.76 | 0.04 | 0.45 | 0.04 | 0.74 | 0.03 | 0.43 | 0.06 | **0.97** | 0.01 | **0.95** | 0.03 |
| 46 | 0.73 | 0.04 | 0.50 | 0.05 | 0.75 | 0.06 | 0.46 | 0.07 | **0.98** | 0.02 | **0.93** | 0.03 |
| 47 | 0.69 | 0.05 | 0.57 | 0.05 | 0.84 | 0.04 | 0.59 | 0.07 | **0.99** | 0.01 | **0.96** | 0.03 |
| 48 | 0.71 | 0.05 | 0.54 | 0.05 | 0.81 | 0.07 | 0.55 | 0.07 | **0.96** | 0.02 | **0.93** | 0.02 |
| 49 | 0.72 | 0.07 | 0.53 | 0.07 | 0.73 | 0.07 | 0.47 | 0.06 | **0.98** | 0.01 | **0.95** | 0.03 |
| 50 | 0.65 | 0.07 | 0.50 | 0.07 | 0.67 | 0.10 | 0.36 | 0.06 | **0.91** | 0.06 | **0.71** | 0.09 |
| 51 | 0.61 | 0.03 | 0.51 | 0.02 | 0.65 | 0.04 | 0.40 | 0.05 | **0.93** | 0.03 | **0.86** | 0.05 |
| 52 | 0.76 | 0.03 | 0.48 | 0.03 | 0.71 | 0.02 | 0.46 | 0.06 | **0.94** | 0.03 | **0.93** | 0.03 |
| 53 | 0.69 | 0.05 | 0.52 | 0.06 | 0.66 | 0.06 | 0.45 | 0.08 | **0.98** | 0.01 | **0.96** | 0.02 |
| 54 | 0.73 | 0.05 | 0.62 | 0.07 | 0.78 | 0.08 | 0.59 | 0.10 | **0.99** | 0.02 | **0.97** | 0.04 |
| 55 | 0.75 | 0.01 | 0.54 | 0.02 | 0.82 | 0.05 | 0.44 | 0.09 | **0.98** | 0.01 | **0.98** | 0.01 |
| 56 | 0.75 | 0.05 | 0.51 | 0.06 | 0.72 | 0.07 | 0.38 | 0.03 | **0.97** | 0.02 | **0.94** | 0.03 |
| 57 | 0.85 | 0.05 | 0.62 | 0.05 | 0.90 | 0.04 | 0.73 | 0.04 | **0.99** | 0.01 | **0.77** | 0.10 |
| 58 | 0.88 | 0.03 | 0.63 | 0.03 | 0.93 | 0.02 | 0.79 | 0.04 | **0.98** | 0.01 | **0.94** | 0.04 |
| 59 | 0.88 | 0.02 | 0.68 | 0.02 | 0.94 | 0.02 | 0.82 | 0.04 | **0.97** | 0.02 | **0.86** | 0.04 |
| 60 | 0.90 | 0.02 | 0.70 | 0.02 | 0.95 | 0.02 | 0.83 | 0.04 | **0.99** | 0.01 | **0.95** | 0.02 |
| 61 | 0.90 | 0.03 | 0.75 | 0.05 | 0.97 | 0.01 | 0.90 | 0.04 | **0.98** | 0.02 | **0.95** | 0.04 |
| 62 | 0.90 | 0.02 | 0.74 | 0.02 | 0.95 | 0.02 | 0.84 | 0.04 | **0.99** | 0.01 | **0.94** | 0.04 |
| 63 | 0.91 | 0.02 | 0.70 | 0.03 | 0.96 | 0.02 | 0.86 | 0.03 | **0.99** | 0.01 | **0.96** | 0.02 |

Table 2: Performance comparison on 63 types of synthetic datasets. Each pair of the score is a TPR score and its standard deviation.

| **Simple** | | **Statistics** | | **Hampel** | | **Filter** | | **Dynamic** | | **Filter** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **bio_outlier** | **Std** | **abnormality** | **Std** | **bio_outlier** | **Std** | **abnormality** | **Std** | **bio_outlier** | **Std** | **abnormality** | **Std** |
| 0.85 | 0.03 | 0.90 | 0.03 | 0.84 | 0.06 | 0.92 | 0.05 | 0.97 | 0.05 | 1.00 | 0.01 |
| 0.85 | 0.03 | 0.90 | 0.05 | 0.87 | 0.06 | 0.88 | 0.06 | 0.98 | 0.02 | 0.99 | 0.01 |
| 0.87 | 0.03 | 0.88 | 0.04 | 0.92 | 0.02 | 0.93 | 0.04 | 0.99 | 0.01 | 1.00 | 0.00 |
| 0.88 | 0.02 | 0.83 | 0.05 | 0.86 | 0.06 | 0.87 | 0.03 | 0.99 | 0.01 | 0.99 | 0.01 |
| 0.86 | 0.06 | 0.84 | 0.03 | 0.86 | 0.05 | 0.83 | 0.03 | 1.00 | 0.00 | 1.00 | 0.00 |
| 0.89 | 0.02 | 0.88 | 0.05 | 0.93 | 0.03 | 0.89 | 0.04 | 0.99 | 0.01 | 0.99 | 0.02 |
| 0.88 | 0.01 | 0.86 | 0.05 | 0.88 | 0.05 | 0.87 | 0.04 | 0.99 | 0.01 | 1.00 | 0.01 |
| 0.83 | 0.03 | 0.91 | 0.05 | 1.00 | 0.00 | 1.00 | 0.00 | 0.97 | 0.02 | 1.00 | 0.01 |
| 0.85 | 0.01 | 0.88 | 0.02 | 0.99 | 0.01 | 1.00 | 0.00 | 0.95 | 0.06 | 1.00 | 0.01 |
| 0.89 | 0.01 | 0.89 | 0.01 | 0.99 | 0.01 | 0.99 | 0.01 | 0.99 | 0.01 | 1.00 | 0.00 |
| 0.86 | 0.02 | 0.89 | 0.03 | 0.99 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 | 0.98 | 0.02 |
| 0.86 | 0.03 | 0.90 | 0.07 | 0.99 | 0.01 | 1.00 | 0.01 | 0.98 | 0.01 | 0.99 | 0.01 |
| 0.88 | 0.02 | 0.90 | 0.03 | 0.99 | 0.01 | 1.00 | 0.00 | 0.99 | 0.01 | 0.99 | 0.01 |
| 0.85 | 0.03 | 0.90 | 0.04 | 1.00 | 0.00 | 1.00 | 0.00 | 0.99 | 0.01 | 1.00 | 0.01 |
| 0.79 | 0.05 | 0.88 | 0.06 | 0.58 | 0.10 | 0.63 | 0.07 | 0.86 | 0.11 | 0.87 | 0.10 |
| 0.82 | 0.03 | 0.88 | 0.04 | 0.66 | 0.07 | 0.73 | 0.09 | 0.96 | 0.04 | 0.95 | 0.08 |
| 0.77 | 0.05 | 0.88 | 0.04 | 0.64 | 0.10 | 0.70 | 0.10 | 0.95 | 0.02 | 1.00 | 0.00 |
| 0.87 | 0.04 | 0.90 | 0.03 | 0.81 | 0.06 | 0.83 | 0.09 | 0.99 | 0.01 | 1.00 | 0.01 |
| 0.87 | 0.05 | 0.87 | 0.07 | 0.71 | 0.11 | 0.82 | 0.08 | 0.98 | 0.02 | 1.00 | 0.00 |
| 0.85 | 0.04 | 0.82 | 0.05 | 0.71 | 0.09 | 0.77 | 0.08 | 0.97 | 0.02 | 0.99 | 0.02 |
| 0.85 | 0.02 | 0.90 | 0.04 | 0.68 | 0.05 | 0.83 | 0.05 | 0.98 | 0.02 | 1.00 | 0.00 |
| 0.81 | 0.03 | 0.95 | 0.04 | 0.88 | 0.03 | 0.94 | 0.04 | 0.95 | 0.03 | 1.00 | 0.01 |
| 0.83 | 0.03 | 0.98 | 0.03 | 0.90 | 0.02 | 0.96 | 0.03 | 0.99 | 0.01 | 1.00 | 0.00 |
| 0.86 | 0.01 | 0.96 | 0.01 | 0.93 | 0.03 | 0.97 | 0.02 | 1.00 | 0.00 | 1.00 | 0.00 |
| 0.89 | 0.02 | 0.94 | 0.04 | 0.96 | 0.01 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 |
| 0.90 | 0.01 | 0.99 | 0.02 | 0.97 | 0.01 | 1.00 | 0.00 | 0.96 | 0.04 | 1.00 | 0.00 |
| 0.89 | 0.01 | 1.00 | 0.00 | 0.97 | 0.02 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 |
| 0.90 | 0.02 | 1.00 | 0.00 | 0.97 | 0.02 | 1.00 | 0.00 | 0.97 | 0.04 | 1.00 | 0.00 |
| 0.71 | 0.01 | 0.80 | 0.02 | 0.85 | 0.04 | 0.91 | 0.02 | 0.97 | 0.02 | 0.95 | 0.08 |
| 0.72 | 0.01 | 0.81 | 0.04 | 0.86 | 0.02 | 0.90 | 0.05 | 0.98 | 0.01 | 1.00 | 0.00 |
| 0.73 | 0.02 | 0.79 | 0.06 | 0.92 | 0.02 | 0.90 | 0.04 | 0.98 | 0.01 | 1.00 | 0.01 |
| 0.73 | 0.02 | 0.82 | 0.07 | 0.94 | 0.02 | 0.95 | 0.03 | 1.00 | 0.00 | 1.00 | 0.00 |
| 0.71 | 0.01 | 0.81 | 0.08 | 0.94 | 0.02 | 0.94 | 0.02 | 0.99 | 0.01 | 0.99 | 0.01 |
| 0.73 | 0.02 | 0.80 | 0.05 | 0.93 | 0.02 | 0.95 | 0.02 | 0.98 | 0.03 | 0.99 | 0.01 |
| 0.72 | 0.02 | 0.74 | 0.04 | 0.93 | 0.01 | 0.92 | 0.02 | 0.98 | 0.01 | 0.99 | 0.01 |
| 0.67 | 0.01 | 0.84 | 0.04 | 0.90 | 0.01 | 0.94 | 0.02 | 0.97 | 0.03 | 0.94 | 0.12 |
| 0.70 | 0.02 | 0.86 | 0.03 | 0.91 | 0.02 | 0.93 | 0.02 | 0.99 | 0.02 | 1.00 | 0.00 |
| 0.70 | 0.03 | 0.93 | 0.03 | 0.94 | 0.03 | 0.94 | 0.02 | 0.99 | 0.02 | 1.00 | 0.00 |
| 0.72 | 0.03 | 0.98 | 0.01 | 0.95 | 0.01 | 0.98 | 0.03 | 0.99 | 0.02 | 1.00 | 0.00 |
| 0.75 | 0.03 | 0.97 | 0.04 | 0.97 | 0.01 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 |
| 0.78 | 0.02 | 0.91 | 0.06 | 0.98 | 0.01 | 1.00 | 0.00 | 0.99 | 0.02 | 1.00 | 0.00 |
| 0.78 | 0.02 | 0.95 | 0.04 | 0.98 | 0.01 | 0.99 | 0.02 | 1.00 | 0.00 | 0.96 | 0.08 |
| 0.61 | 0.03 | 0.70 | 0.03 | 0.56 | 0.08 | 0.60 | 0.09 | 0.91 | 0.02 | 0.97 | 0.02 |
| 0.65 | 0.04 | 0.71 | 0.03 | 0.60 | 0.04 | 0.66 | 0.13 | 0.94 | 0.02 | 0.95 | 0.08 |
| 0.61 | 0.02 | 0.75 | 0.05 | 0.63 | 0.03 | 0.70 | 0.04 | 0.96 | 0.01 | 0.98 | 0.04 |
| 0.67 | 0.03 | 0.69 | 0.07 | 0.66 | 0.03 | 0.67 | 0.06 | 0.94 | 0.02 | 1.00 | 0.01 |
| 0.64 | 0.04 | 0.83 | 0.08 | 0.76 | 0.06 | 0.75 | 0.02 | 0.96 | 0.03 | 1.00 | 0.00 |
| 0.67 | 0.05 | 0.77 | 0.09 | 0.75 | 0.07 | 0.72 | 0.05 | 0.94 | 0.02 | 0.99 | 0.02 |
| 0.65 | 0.03 | 0.77 | 0.05 | 0.65 | 0.06 | 0.70 | 0.02 | 0.97 | 0.02 | 0.98 | 0.02 |

Table. 2 (cont'd)

| **Simple** | | **Statistics** | | **Hampel** | | **Filter** | | **Dynamic** | | **Filter** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **bio_outlier** | **Std** | **abnormality** | **Std** | **bio_outlier** | **Std** | **abnormality** | **Std** | **bio_outlier** | **Std** | **abnormality** | **Std** |
| 0.59 | 0.03 | 0.77 | 0.02 | 0.61 | 0.07 | 0.59 | 0.04 | 0.84 | 0.03 | 0.80 | 0.13 |
| 0.61 | 0.03 | 0.75 | 0.02 | 0.60 | 0.03 | 0.68 | 0.03 | 0.90 | 0.04 | 0.95 | 0.08 |
| 0.62 | 0.03 | 0.75 | 0.02 | 0.66 | 0.04 | 0.67 | 0.07 | 0.94 | 0.03 | 0.99 | 0.01 |
| 0.62 | 0.02 | 0.79 | 0.05 | 0.60 | 0.06 | 0.72 | 0.06 | 0.98 | 0.01 | 0.98 | 0.02 |
| 0.73 | 0.05 | 0.79 | 0.12 | 0.74 | 0.05 | 0.78 | 0.09 | 0.98 | 0.03 | 0.99 | 0.02 |
| 0.67 | 0.05 | 0.78 | 0.07 | 0.68 | 0.04 | 0.66 | 0.09 | 0.98 | 0.01 | 1.00 | 0.00 |
| 0.69 | 0.05 | 0.72 | 0.08 | 0.60 | 0.04 | 0.67 | 0.04 | 0.95 | 0.02 | 0.99 | 0.03 |
| 0.79 | 0.03 | 0.85 | 0.04 | 0.86 | 0.02 | 0.90 | 0.02 | 0.88 | 0.07 | 0.89 | 0.06 |
| 0.77 | 0.03 | 0.90 | 0.06 | 0.90 | 0.02 | 0.93 | 0.03 | 0.96 | 0.02 | 0.99 | 0.02 |
| 0.84 | 0.03 | 0.82 | 0.01 | 0.91 | 0.03 | 0.94 | 0.04 | 0.94 | 0.03 | 0.90 | 0.09 |
| 0.84 | 0.01 | 0.87 | 0.03 | 0.92 | 0.02 | 0.93 | 0.03 | 0.98 | 0.01 | 0.98 | 0.02 |
| 0.87 | 0.04 | 0.90 | 0.02 | 0.95 | 0.02 | 0.97 | 0.02 | 0.97 | 0.02 | 1.00 | 0.00 |
| 0.88 | 0.02 | 0.89 | 0.02 | 0.92 | 0.03 | 0.95 | 0.03 | 0.97 | 0.02 | 0.99 | 0.01 |
| 0.85 | 0.02 | 0.89 | 0.02 | 0.93 | 0.01 | 0.95 | 0.02 | 0.97 | 0.02 | 0.99 | 0.01 |

# BIBLIOGRAPHY

# BIBLIOGRAPHY


[1] I Ajjawi, Y Lu, LJ Savage, SM Bell, and RL Last. Large-scale reverse genetics in arabidopsis: Case studies from the chloroplast 2010 project. *Plant Physiol*, 152(2):529–540, 2010.

[2] I Ajjawi, Y Lu, LJ Savage, SM Bell, and RL Last. Large-scale reverse genetics in arabidopsis: case studies from the chloroplast 2010 project. *Plant physiology*, 152(2):529–540, 2010.

[3] JM Alonso, AN Stepanova, TJ Leisse, et al. Genome-wide insertional mutagenesis of arabidopsis thaliana. *Science*, 301(5633):653–657, 2003.

[4] NS Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *Am Stat*, 46(3):175–185, 1992.

[5] P Baldi, S Brunak, Y Chauvin, C Andersen, and H Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[6] DM Bates and DG Watts. *Nonlinear regression: iterative estimation and linear approximations*. Wiley Online Library, 1988.

[7] CM Bishop. *Pattern recognition and machine learning*. springer, 2006.

[8] William M Bolstad. *Introduction to Bayesian statistics*. John Wiley & Sons, 2013.

[9] J Bonner. The upper limit of crop yield this classical problem may be analyzed as one of the photosynthetic efficiency of plants in arrays. *Science*, 137(3523):11–15, 1962.

[10] CA Cameron and FAG Windmeijer. An r-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, 77(2):329–342, 1997.

[11] TC Chou and P TaLaLay. Generalized equations for the analysis of inhibitions of michaelis-menten and higher-order kinetic systems with two or more mutually exclusive and nonexclusive inhibitors. *European Journal of Biochemistry*, 115(1):207–216, 1981.

[12] F Chu, Y Wang, DS Parker, and C Zaniolo. Data cleaning using belief propagation. *IQ2S*, pages 99–104, 2005.

[13] X Chu, IF Ilyas, and P Papotti. Holistic data cleaning: Putting violations into context. *ICDE*, pages 458–469, 2013.

[14] JA Cruz, LJ Savage, R Zegarac, W Kovac, C Hall, J Chen, and DM Kramer. Dynamic environmental photosynthetic imaging (depi) reveals emergent phenotypes related to the environmental responses of photosynthesis. *Nat Biotech*, in revision, 2014.

[15] F de A. Lobo, MP de Barros, HJ Dalmagro, et al. Fitting net photosynthetic light-response curves with microsoft excel Â¡Âª a critical look at the models. *Photosynthetica*, 51(3):445–456, 2013.

[16] JE Dowd and DS Riggs. A comparison of estimates of michaelis-menten kinetic constants from various linear transformations. *Journal of biological Chemistry*, 240(2):863–869, 1965.

[17] A Ebaid, A Elmagarmid, IF Ilyas, M Ouzzani, JA Quiane-Ruiz, N Tang, and S Yin. Nadeef: a generalized data cleaning system. *VLDB Endowment*, 6(12):1218–1221, 2013.

[18] P Eilers and J Peeters. A model for the relationship between light intensity and the rate of photosynthesis in phytoplankton. *Ecological modelling*, 42(3):199–215, 1988.

[19] F Fiorani and U Schurr. Future scenarios for plant phenotyping. *Annu Rev Plant Biol*, 64:267–291, 2013.

[20] DA Freedman. *Statistical models: theory and practice*. Cambridge University Press, 2009.

[21] Q Gao, E Ostendorf, JA Cruz, R Jin, DM Kramer, and J Chen. Inter-functional analysis of high-throughput phenotype data by nonparametric clustering and its application in photosynthesis. *Bioinformatics*, btv515, 2015.

[22] A Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

[23] Govindjee, Beatty JT, Gest H, and Allen JF. *Discoveries in Photosynthesis*. Springer, 2005.

[24] JM Green, H Appel, EM Rehrig, J Harnsomburana, JF Chang, P Balint-Kurti, and CR Shyu. Phenophyte: a flexible affordable method to quantify 2d phenotypes from imagery. *Plant Methods*, 8(1):1–12, 2012.

[25] DK Großkinsky, J Svensgaard, S Christensen, and T Roitsch. Plant phenomics and the need for physiological phenotyping across scales to narrow the genotype-to-phenotype knowledge gap. *Journal of experimental botany*, 66(18):5429–5440, 2015.

[26] MR Gupta, EK Garcia, and E Chin. Adaptive local linear regression with application to printer color management. *Image Processing, IEEE Transactions on*, 17(6):936–945, 2008.

[27] KG Herbert, NH Gehani, WH Piel, JTL Wang, and CH Wu. Bio-ajax: an extensible framework for biological data cleaning. *ACM SIGMOD Record*, 33(2):51–57, 2004.

[28] PW Holland and RE Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-Theory and Methods*, 6(9):813–827, 1977.

[29] D Houle, DR Govindaraju, and S Omholt. Phenomics: the next challenge. *Nat Rev Genet*, 11(12):855–866, 2010.

[30] AD Jassby and T Platt. Mathematical formulation of the relationship between photosynthesis and light for phytoplankton. *American Society of Limnology and Oceanography*, pages 540–547, 1976.

[31] JLY Koh, ML Lee, W Hsu, and KT Lam. Correlation-based detection of attribute outliers. In *Advances in Databases: Concepts, Systems and Applications*, pages 164–175. 2007.

[32] DM Kramer and JR Evans. The importance of energy balance in improving photosynthetic productivity. *Plant physiol*, 155(1):70–78, 2011.

[33] M Kutsukake, XY Meng, N Katayama, N Nikoh, H Shibao, and T Fukatsu. An insect-induced novel plant phenotype for sustaining social life in a closed system. *Nature communications*, 3:1187, 2012.

[34] H Lambers, FS Chapin III, and TL Pons. Mechanisms of photosynthetic oxygen evolution and fundamental hypotheses of photosynthesis. In *Plant Physiological Ecology*, pages 26–47. Springer, 2008.

[35] K Levenberg. A method for the solution of certain non–linear problems in least squares. (2):164–168, 1944.

[36] XP Li et al. A pigment-binding protein essential for regulation of photosynthetic light harvesting. *Nature*, 403(6768):391–395, 2000.

[37] SP Long and JE Hällgren. Measurement of co2 assimilation by plants in the field and the laboratory. In D.O. Hall, J.M.O. Scurlock, H.R. Bolhàr-Nordenkampf, et al., editors, *Photosynthesis and Production in a Changing Environment*, pages 129–167. Chapman and Hall, 1993.

[38] HL MacIntyre, TM Kana, T Anning, and RJ Geider. Photoacclimation of photosynthesis irradiance response curves and photosynthetic pigments in microalgae and cyanobacteria. *J Phycol*, 38(1):17–38, 2002.

[39] JI Maletic and A Marcus. Data cleansing: Beyond integrity analysis. In *IQ*, pages 200–209, 2000.

[40] C Mayfield, J Neville, and S Prabhakar. Eracer: a database approach for statistical inference and data cleaning. *ACM SIGMOD*, pages 75–86, 2010.

[41] G McLachlan. *Discriminant analysis and statistical pattern recognition*, volume 544. John Wiley & Sons, 2004.

[42] L Menten and MI Michaelis. Die kinetik der invertinwirkung. *Biochem Z*, 49:333–369, 1913.

[43] H Motulsky and A Christopoulos. *Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting.* Oxford University Press, 2004.

[44] H Muller and JC Freytag. *Problems, methods, and challenges in comprehensive data cleansing.* Professoren des Inst. Fur Informatik, 2005.

[45] P. Muller, X. Li, and K.K. Niyogi. Non-photochemical quenching. a response to excess light energy. *Plant Phisiol*, 125(4):1558–66, 2001.

[46] AB Nicotra, OK Atkin, SP Bonser, AM Davidson, EJ Finnegan, U Mathesius, P Poot, MD Purugganan, CL Richards, F Valladares, et al. Plant phenotypic plasticity in a changing climate. *Trends in plant science*, 15(12):684–692, 2010.

[47] E. Ögren and J. R. Evans. Photosynthetic light-response curves. *Planta*, 189(2):182–190, 1993.

[48] RK Pearson. Outliers in process modeling and identification. *IEEE T Contr Syst T*, 10(1):55–63, 2002.

[49] RK Pearson. *Mining imperfect data: Dealing with contamination and incomplete records*. Siam, 2005.

[50] TD Price, A Qvarnström, and DE Irwin. The role of phenotypic plasticity in driving genetic evolution. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(1523):1433–1440, 2003.

[51] CR Rao. The utilization of multiple measurements in problems of biological classification. *J R Stat Soc*, 10(4):159–203, 1948.

[52] U Rascher, S Blossfeld, F Fiorani, et al. Non-invasive approaches for phenotyping of enhanced performance traits in bean. *Funct Plant Biol*, 38(12):968–983, 2011.

[53] D Reynolds. Gaussian mixture models. *Encyclopedia of Biometrics*, pages 659–663, 2009.

[54] G Seber and C Wild. *Nonlinear Regression*. Wiley-Interscience, 2003.

[55] J Serôdio and J Lavaud. A model for describing the light response of the nonphotochemical quenching of chlorophyll fluorescence. *Photosynthesis research*, 108(1):61–76, 2011.

[56] M Shanahan. Perception as abduction: Turning sensor data into meaningful representation. *Cognitive Sci*, 29(1):103–134, 2005.

[57] H Sohn, DW Allen, K Worden, and CR Farrar. Structural damage classification using extreme value statistics. *J dyn syst-t asme*, 127(1):125–132, 2005.

[58] S Subramaniam, T Palpanas, D Papadopoulos, V Kalogeraki, and D Gunopulos. Online outlier detection in sensor data using non-parametric models. *VLDB*, pages 187–198, 2006.

[59] R Subramanian, EP Spalding, and NJ Ferrier. A high throughput robot system for machine vision based plant phenotype studies. *Machine Vision and Applications*, 24(3):619–636, 2013.

[60] K Takizawa, JA Cruz, A Kanazawa, and DM Kramer. The thylakoid proton motive force in vivo. quantitative, non-invasive probes, energetics, and regulatory consequences of light-induced pmf. *Biochim. Biophys*, 1767(10):1233–44, 2007.

[61] OL Tessmer, Y Jiao, JA Cruz, DM Kramer, and J Chen. Functional approach to high-throughput plant growth analysis. *BMC Syst Biol*, 7(Suppl 6):S17, 2013.

[62] OL Tessmer, Y Jiao, JA Cruz, DM Kramer, and J Chen. Functional approach to high-throughput plant growth analysis. *BMC System Biology*, 7(Suppl 6):S17, 2013.

[63] C Thrampoulidis, S Oymak, and B Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Proceedings of The 28th Conference on Learning Theory*, pages 1683–1709, 2015.

[64] J Vlasblom, K Zuberi, H Rodriguez, R Arnold, A Gagarinova, V Deineko, A Kumar, E Leung, K Rizzolo, B Samanfar, et al. Novel function discovery with genemania: a new integrated resource for gene function prediction in escherichia coli. *Bioinformatics*, page btu671, 2014.

[65] S Von Caemmerer and GD Farquhar. Some relationships between the biochemistry of photosynthesis and the gas exchange of leaves. *Planta*, 153(4):376–387, 1981.

[66] A Walter, F Liebisch, and A Hund. Plant phenotyping: from bean weighing to image analysis. *Plant methods*, 11(1):14, 2015.

[67] Andrew R. Webb. *FisherÂ¡Â¯s criterion Â¨C linear discriminant analysis*. Wiley, 2002.

[68] ML Wong, C Dong, V Andreev, M Arcos-Burgos, and J Licinio. Prediction of susceptibility to major depression by a model of interactions of multiple functional genetic variants and environmental factors. *Molecular psychiatry*, 17(6):624–633, 2012.

[69] L. Xu, JA. Cruz, L. Savage, DM. Kramer, and J. Chen. Plant photosynthesis phenomics data quality control. *Bioinformatics*, 31(11):1796–1804, 2015.

[70] J Yang and J Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186, 2011.

[71] X Yin, X Liu, J Chen, and DM Kramer. Multi-leaf alignment from fluorescence plant images. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 437–444, 2014.

[72] X Yin, X Liu, J Chen, and DM Kramer. Multi-leaf tracking from fluorescence plant videos. In *Image Processing(ICIP), 2014 IEEE International Conference on*, pages 408–412, 2014.

[73] Y Zeinalov. Mechanisms of photosynthetic oxygen evolution and fundamental hypotheses of photosynthesis.

[74] X Zhu, Stephen P. Long, and Donald R. Ort. Improving photosynthetic efficiency for greater yield. *Annu Rev Plant Biol*, 61:235–261, 2010.