RESPIRATORY DISEASE DIAGNOSIS FOR DOLPHIN USING BREATH DATA

A Thesis

by

CHINH THI TUYET LAI

BS, Texas A&M University-Corpus Christi, 2015

Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in

MATHEMATICS

Texas A&M University-Corpus Christi
Corpus Christi, Texas

August 2017

August 2017

RESPIRATORY DISEASE DIAGNOSIS FOR DOLPHIN USING BREATH DATA

A Thesis

by

CHINH THI TUYET LAI

This thesis meets the standards for scope and quality of
Texas A&M University-Corpus Christi and is hereby approved.

LEI JIN, PhD
Chair

JOSE GUARDIOLA, PhD
Committee Member

BLAIR STERBA-BOATWRIGHT, PhD
Committee Member

August 2017

ABSTRACT

Respiratory disease in marine mammals evokes strong public attention as well as worthwhile scientific interest. Traditional methods for animal disease diagnosis include blood test, ultrasound, and computed tomography scan. These methods require invasive equipment to perform, and cannot be applied to free-swimming animals.

Breath data, the measurement of lung inflow and outflow while breathing, can be collected from free-swimming animals in a non-invasive way, and so is less stressful for distressed animals. However, because of new features in the data, new statistical methods are required for the breath data analysis. In this thesis, we investigate one potential method for analyzing breath data. Our method begins by decomposing a raw dataset containing a sequence of breath cycles into a set of individual breath cycles. Incomplete cycles are removed from the dataset. In this research, we consider an entire breath cycle to be one unit of observation. Starting and ending points of breath cycles can be difficult to determine, and cause a large amount of variation in size and shape of breath curves. To reduce cycle to cycle variability, we apply curve registration to synchronize a set of breath cycles. Breath cycles are described using magnitude information and geometric shape information. We propose three shape models, namely, simple oval model, quadratic spline model, and piecewise linear model. Furthermore, principal component analysis is applied to the magnitude/shape descriptors to obtain main features of breath cycles. Criteria for disease diagnosis are developed by identifying key differences among these main features between healthy and unhealthy animals. The proposed methods were applied to check if two testing animals are diseased or not. The results were consistent with the status of both animals.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLES PAGE

CHAPTER I: INTRODUCTION

The growing number of marine mammals suffering morbidity and mortality from respiratory disease implies a growing need for diagnosing the disease [23]. In this chapter, we briefly introduce traditional and modern methods for diagnosing respiratory disease with respect to their advantages and disadvantages. Due to benefits from the analysis of breath data, it becomes the main focus of this study. The use of functional data analysis with computational statistics is introduced to provide tools to analyze breath data. The chapter ends with a review of principle component analysis, which is used to extract main features of breath cycles.

1.1 Traditional methods of respiratory disease diagnosis for dolphin

Typical medical imaging methods, including ultrasound and computed tomography (CT) scan, enable veterinarian to visualize and determine health condition of lung system. In addition, the analysis of blood sample, and respiration cycle help to examine the working ability and health condition of lung system. In this section, we briefly describe each individual diagnostic approach with respect to its advantages and disadvantages. Finally, we explain why the analysis of breath cycles is chosen as the main method in our research.

To draw comprehensive picture of lung system, ultrasound uses high-frequency sound wave, and CT scan uses high-energy electromagnetic wave. Veterinarian then can visualize and identify abnormal condition that can be a sign of lung disease. Besides that, blood test is taken in medical laboratory, where technician extracts blood sample and perform multiple tests to analyze each component included in the blood sample. Abnormal level of certain components in the blood sample can be a sign of certain diseases [4]. For example, the present of miR, an intronic microRNA, can be a sign of lung cancer [11].

Although blood test, ultrasound, and CT scan are common methods in animal disease diagnosis, these methods are difficult to perform on free-swimming animals living in the aquatic envi-

ronment. Blood test requires extracting blood sample from dolphin, which requires special care in monitoring the health condition of the dolphin before and after taking blood. In addition, experienced staff must be trained in the extraction method. As for ultrasound and CT scan, these require dolphin to be placed in the relevant machines to take a series of pictures, which again involves special care and the need for trained and experienced technician.

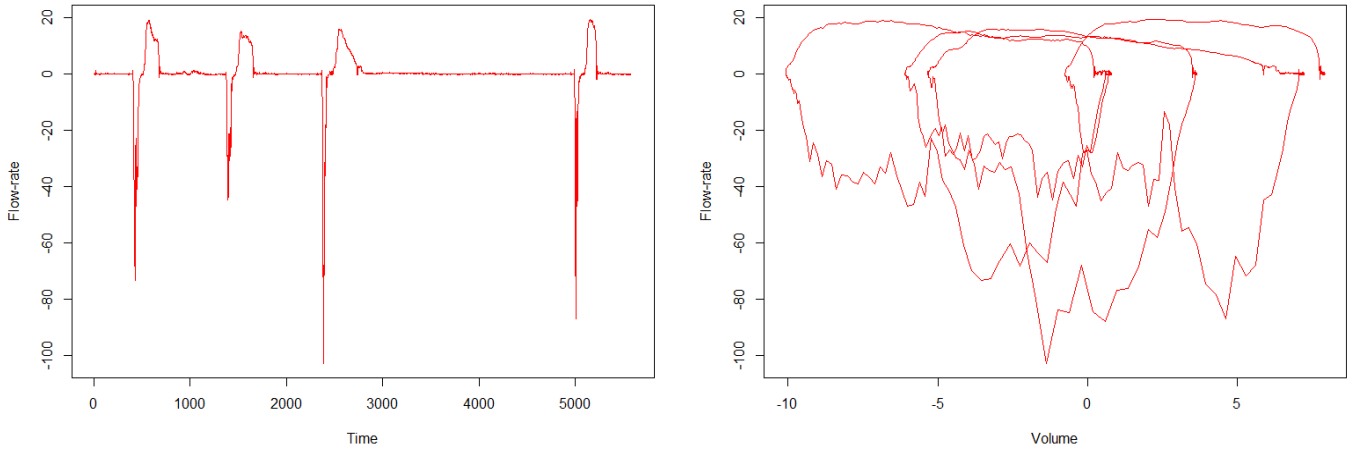
Compared to the traditional methods, the analysis of breath data is a non-invasive way for diagnosing respiratory disease in dolphin. Respiration cycles of dolphin are measured by a combined flow-meter (pneumotachometer), making it easy to collect breath data. Then, functional data analysis is used to analyze breath data. The findings and results of the study will help researchers to diagnose health status of marine mammals, which can be used in studying marine biology.

1.2 Description of breath data

Our breath data were provided by Dr. Andreas Fahlman, and were collected from different dolphins around the world. Breath data were measured on both healthy and diseased dolphins. Our analysis will focus on two specific measurements, volume and flow-rate. Volume, measured in liters, is the amount of air in the lung at different times as a dolphin exhales and inhales. Flow-rate, measured in liters per seconds, is the change in volume per unit time. Because measurements are taken at the constant frequency rate, for convenience, we just use the temporal order to denote the time information.

In this study, we define a breath cycle to begin with exhalation and end with inhalation. When the dolphin begins to exhale, the flow-rate of a breath cycle starts decreasing from zero. The flow-rate declines to a minimum value before increasing to a maximum value, and then returns to zero again. Consequently, the sign of flow-rate changes from negative to positive as the respiration cycle of the dolphin changes from exhalation to inhalation. Figure 1.1 shows different views of breath cycles. In particular, Figure 1.1a shows breath cycles with respect to flow-rate versus time, which form fluctuating curve moving in forwarding direction. The intervals of fluctuation in Figure 1.1a represent the time intervals when the dolphin exhales and inhales. Besides that, Figure 1.1b shows

the flow-rate of breath cycles plotted against volume, which results in closed and asymmetrical curves. In addition, we define half-up cycles are where flow-rate is greater than or equal to zero, and half-down cycles are where flow-rate is less than zero.



(a) Flow-rate versus time

(b) Flow-rate against volume

Figure 1.1: Breath curves

1.3 Introduction to functional data analysis

Functional data appear in many forms, such as function, curve, surface, or image. In general, a functional data can be represented as a set of independent functions $x_i = x_i(t_i)$ for $i = 1...N$ [14]. The variable t_i usually represents time while x_i represents some other variables.

Functional data analysis (FDA) is a unified collection of statistical techniques to analyze functional data [15]. The history of FDA began in 1958 when Rao studied the growing process of an organism by observing and analyzing its growth curve [16]. He successfully constructed a simple stochastic model of the growth curves. His findings/results helped to examine different growth curves under various conditions. Even though FDA has been involved in the statistical treatment of data since that time, the term ”functional data analysis” has become well-known only since 1982, when Ramsay developed a data analytic technique that was used to study the outcomes of random variables [13]. He expressed the problem straightforwardly in the manner of functional analysis,

in which a datum was described as a continuous function of variable time observed over a specific interval.

The fundamental aim of FDA is to determine a relationship between the input x_i and a targeted function y_i through a map f, which is described as

$$y_i = f(x_i) = f(x_i(t_i))$$

In this study, the breath data is expressed as the function of N, where volume $V = V(N)$ and flow-rate $F = F(N)$. The functional analysis is performed to find a map f, which is defined as

$$F = F(N) = f(V) = f(V(N))$$

Because original breath cycles from raw dataset contain a large amount of variation that would interfere with any synthesis of a common pattern, the analysis of breath data without any modification may lead to certain difficulties and meaningless results. A technique to overcome these problems is curve registration. Curve registration is a statistical technique that synchronizes a set of functional data to create a new dataset in which each datum shares similarity in form and magnitude [14]. The history of curve registration may be traced back to 1978, when Sakoe and Chiba published a method named wrapping function [19]. In their research, the authors collected voice data from different people to create a spoken word recognition model. Since the speaking rate is varied from person to person, they applied several adjustments to convert the dataset into a common range which was suitable for further analysis. The term ”curve registration” was first used by Silverman in 1995 [21]. In his study, Silverman modified previously collected data with many perturbations to functional data, and created a more general procedure to adjust Canadian temperature data into a pre-defined interval and shape.

In 1982, Ramsay introduced one of the most common curve registration techniques, which is mark registration [13]. According to the technique, particular points in a dataset which have special property are detected as marks. In this paper, examples of marks include maxima/minima, and starting/ending points of a cycle.

1.4 Principal component analysis

To diagnose the health condition of dolphin, differences between distributions of scores for healthy and diseased dolphins need to be revealed. Due to the nature of functional data, the breath data are very high dimensional, so it is crucial to reduce the number of dimensions. Principle component analysis (PCA) is a widely-used mathematical algorithm for dimension reduction [17]. The history of PCA may be traced back to 1901, when Pearson used the principal axis theorem in mechanics [12]. In his article, he introduced the method of fitting a vector subspace to a multivariate dataset. His findings/results became the fundamental idea of PCA, in which an input sample data is linearly transformed into a lower-dimensional data that still retain a large enough amount of variation [1].

PCA accomplishes the dimensional reduction process by determining directions, called components, along which the amount of variation in sample data is at maxima [6]. Components with high variances are called principal components, which names the method. As a result, by using principal components, sample data can be described by a set of few numbers, which can be displayed by plot for visualizing and examining purposes. In the study, we apply PCA to obtain principal components of the breath data, which helps to define criteria for respiratory disease diagnosis in dolphin.

CHAPTER II: DECOMPOSING TIME SERIES INTO INDIVIDUAL BREATH CYCLES

In this chapter, we describe how raw breath data time series of raw breath data are decomposed into individual breath cycles. The process is described in the flowchart shown in Figure 2.2.

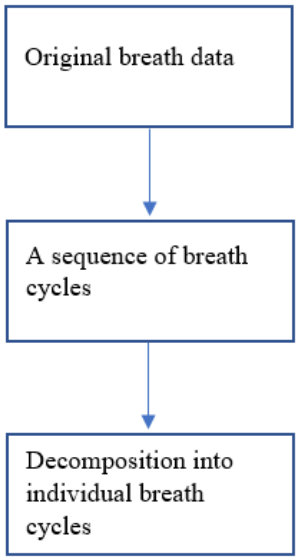


Figure 2.2: Decomposing time series of raw breath data into individual breath cycles

First, original breath data contains a sequence of breath cycles connected by flat segments as may be seen in Figure 2.3. Because respiration cycles occur when flow-rate is changing over time, these steady segments are not connected with breath cycles and may be eliminated from the dataset.

Since breath cycles are defined to begin with exhalation and end with inhalation, the algorithm identifies the start of a breath cycle when the flow-rate begins decreasing continuously from zero, and the end when the flow-rate decreases from a positive value to zero. As a result, incomplete cycles, including either exhalation or inhalation, can be removed from the dataset. Figure 2.4 shows a breath cycle, which is successfully selected from the sequence of cycles.

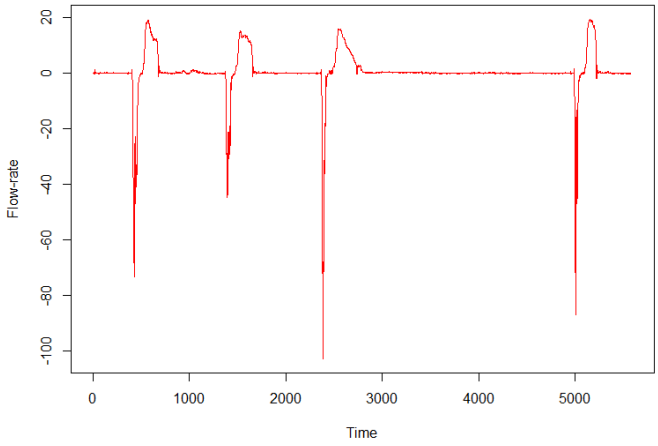
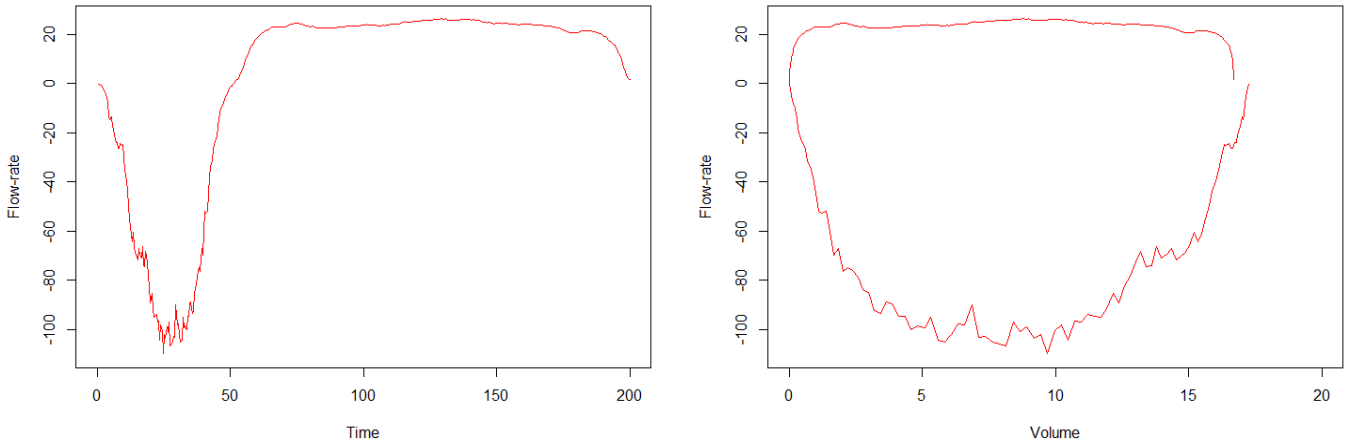


Figure 2.3: A sequence of breath cycles



(a) Flow-rate versus time

(b) Flow-rate against volume

Figure 2.4: A breath cycle

CHAPTER III: METHODS

Due to the functional nature of breath cycles, we use FDA to analyze the breath data as described in the flowchart shown in Figure 3.5. Firstly, we apply curve registration, where multiple adjustments are applied to reduce variability in the breath cycles. After different geometric models are constructed to describe magnitude/shape features of the breath cycles, PCA is used to obtain main features from magnitude/shape descriptors. Then, distributions of scores from the breath data of healthy animals are constructed. Finally, criteria for disease diagnosis are defined to detect health condition of dolphin.

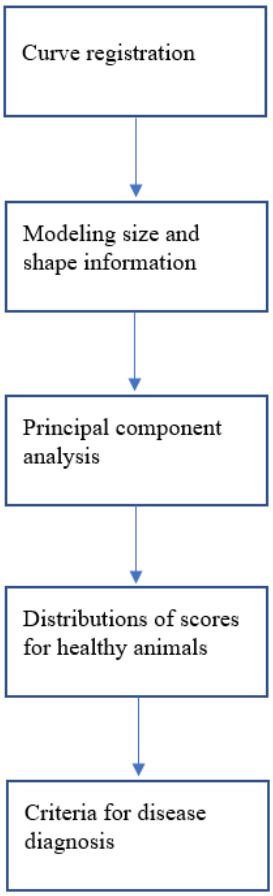


Figure 3.5: Methods in FDA

3.1 Curve registration

Consulting Figure 1.1b, we can that the raw breath curves contain substantial imperfections that would interfere with any synthesis of a common pattern. These imperfections include the differences in starting/ending points and sizes of the breath cycles, and the local intrasubject variability of the breath curves. The purpose of curve registration is to remedy these imperfections without losing any important information. To begin the registration process, we apply moving averages to moderate the intrasubject variability of breath cycles. Moving averages replace a given dataset with sets of averages of consecutive values in that dataset [3]. Since the half-up of the breath curve typically has less variation than the half-down cycle, two different equations are used to describe how moving averages are applied in each half cycle. To illustrate, let $\{X_1, X_2, \ldots X_n\}$ be a set of breath data measurements before moving averages. Then, $\{Y_1, Y_2, \ldots, Y_{n-1}\}$ and $\{Z_1, Z_2, \ldots, Z_{n-2}\}$ describe two new sets of breath data measurements after moving average, which are computed by

$$Y_{i-1} = \frac{X_{i-1} + X_i}{2}, 2 \leq \text{i} \leq \text{n} \tag{3.1}$$

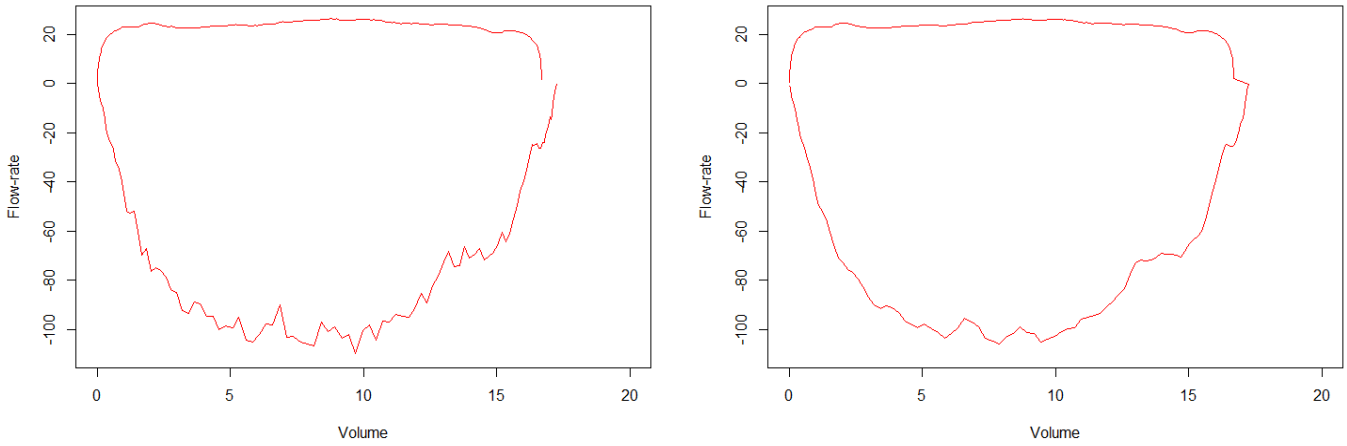
$$Z_{i-2} = \frac{X_{i-2} + X_{i-1} + X_i}{3}, 3 \leq \text{i} \leq \text{n} \tag{3.2}$$

where Y_{i-1} is used to compute the new set of breath data for the top half, and Z_{i-2} is used to compute the new set of breath data for the down half. Figure 3.6 shows that after moving average, the adjusted curve has less local intrasubject variation than the original curve.

Figure 1.1b shows that the magnitude of the breath cycles are different. Our next step is to resize inner area of the breath cycles. Each half of the breath cycle is rescaled separately. We use ratios r_1 and r_3 to rescale the half-up cycle to have maximum flow-rate at 100, and ending point at (200,0). We use ratios r_2 and r_4 to rescale the half-down cycle to have minimum flow-rate at -100, and ending point at (200,0). The equations for computing the ratios are described as

$$r_1 = \frac{100}{F_{mu}} \tag{3.3}$$

$$r_2 = -\frac{100}{F_{md}} \tag{3.4}$$



(a) Before moving averages

(b) After moving averages

Figure 3.6: A breath cycle before and after moving averages

$$r_3 = \frac{200}{V_{mu}} \tag{3.5}$$

$$r_4 = \frac{200}{V_{md}} \tag{3.6}$$

where F_{mu} is maximum flow-rate of the half-up cycle, and F_{md} is minimum flow-rate of the half-down cycle. The quantity volume is defined in a similar fashion. After being resized, maximum flow-rate is equal to 100, minimum flow-rate is equal to -100, and maximum volume is equal to 200. Figure 3.7 shows the breath cycles after curve registration.

3.2 Modeling size and shape information

In this section, we describe three models that provide simplified descriptions of the breath cycles. The first, a simple oval model, uses only the magnitude information of the breath cycles. The remaining two models, a quadratic spline model and a piecewise linear model, augment magnitude information with mark registration of the half-up breath cycles.

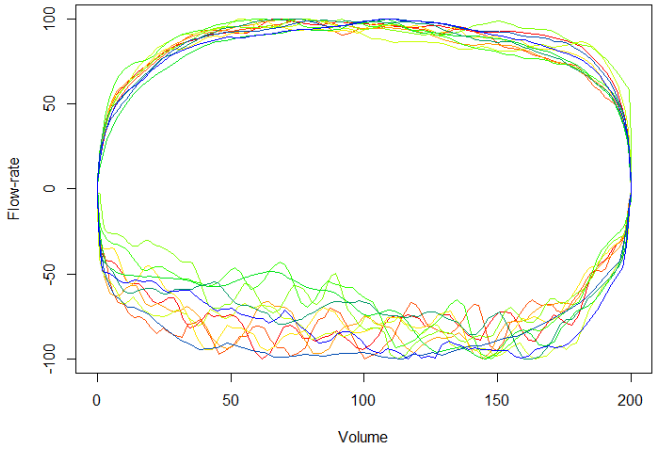


Figure 3.7: Breath cycles after curve registration

3.2.1 Simple oval model

The first geometric model constructed in this study is a simple oval model. To represent the model, a vector of five components is used, which includes area a_e, and the ratios (r_1, r_2, r_3, and r_4). In detail, the ratios r_1, r_2, r_3, and r_4 are used to rescale the magnitude of the breath cycles stay. After being rescaling, the shape of the breath cycles look relatively similar to oval shape, which names the method. Besides that, area a_e, represents inner area bounded by the curve of flow-rate of the breath cycles plotted against volume, is used to describe the magnitude of the entire breath cycles before moving averages and rescaling. The area is calculated by using the trapezoid sum, which is described as [8]

$$area = \frac{US + LS}{2} \tag{3.7}$$

where the upper sum $US = \sum_{i=2}^{i=n} |F_i| \delta V$, the lower sum $LS = \sum_{i=1}^{i=n-1} |F_i| \delta V$, and $\delta V = V_{i+1} - V_i$.

Some calculation results for components of the simple oval model for a sample of 9 breath cycles are shown in Table 3.1.

Cycle	1	2	3	4	5	6	7	8	9
a_e	204.961	163.215	191.894	244.183	233.906	164.690	151.694	70.875	195.070
r_1	4.692	4.878	4.951	4.749	4.970	6.158	6.015	6.085	4.916
r_2	5.531	6.436	5.182	4.561	6.293	5.490	4.624	9.422	5.603
r_3	31.107	35.327	32.678	27.387	25.078	31.321	36.282	56.925	30.077
r_4	31.001	35.318	32.440	27.545	25.059	31.109	35.947	58.388	30.240

Table 3.1: Components of the simple oval model for several breath cycles

3.2.2 Quadratic spline model

The second geometric model constructed in this study is a quadratic spline model. To represent the model, a vector of ten components is used. Firstly, to summarize the magnitude of the breath cycles, we use the inner area a_u of the half-up cycles before curve registration, and the ratios (r_1 and r_3). The fourth component of the vector is x_m, the volume after curve registration at while the maximum re-scaled flow rate of 100 occurs. Secondly, we use a_1, b_1, a_2, and b_2 to summarize the geometric shape of half-up cycles after curve registration. The name "quadratic spline", itself, represents that quadratic equation is used to construct the model. The system of equations below describes how quadratic equation is applied to construct the model.

$$\begin{cases} f_1(x) &= a_1 \times x^2 + b_1 \times x + c_1, \text{when x} \leq x_m \\ f_2(x) &= a_2 \times x^2 + b_2 \times x + c_2, \text{when x} \geq x_m \\ f_1'(x_m) &= f_2'(x_m) \end{cases} \tag{3.8}$$

where a_1 and b_1 are the coefficients of the first quadratic equation, which passes through the origin and maxima at (x_m, 100) after curve registration; and a_2 and b_2 are the coefficients of the second quadratic equation, which passes through the maxima and ending point at (200, 0) after curve registration. Additionally, at the maximum rescaled flow-rate, the first derivative of the two quadratic equations are equal to each other.

The last two components of the vector representing the quadratic spline model are R_1 and R_2, which summarize deviation from the quadratic spline model. The equations for computing R_1 and R_2 are described as

$$R_1 = y_2 - y_1 \tag{3.9}$$

$$R_2 = y_4 - y_3 \tag{3.10}$$

where x_1 is the midpoint of the interval [0, x_m], y_2 is the positive flow-rate at x_1, and y_1 is the predicted flow-rate for the quadratic spline at x_1. The quantity R_2 is computed in a similar fashion on the interval [x_m, 200].

Some calculation results for components of the quadratic spline model for a sample of 9 breath cycles are shown in Table 3.2. Figure 3.8 shows an example of quadratic spline model.

Cycle	1	2	3	4	5	6	7	8	9
a_u	118.421	95.891	102.516	128.721	134.847	90.893	76.869	46.454	113.084
r_1	4.692	4.878	4.951	4.749	4.970	6.158	6.015	6.085	4.916
r_3	31.107	35.327	32.678	27.387	25.078	31.321	36.282	56.925	30.077
x_m	68.489	78.312	71.445	62.884	55.164	76.806	59.341	108.827	76.281
a_1	-0.021	-0.016	-0.019	-0.025	-0.032	-0.017	-0.028	-0.008	-0.017
b_1	2.891	2.505	2.683	3.091	3.559	2.573	3.288	1.828	2.565
a_2	-0.006	-0.006	-0.006	-0.005	-0.005	-0.005	-0.005	-0.012	-0.006
b_2	0.757	0.987	0.814	0.615	0.518	0.724	0.564	2.511	0.954
R_1	20.507	23.209	19.471	19.482	20.045	22.841	18.971	26.767	19.859
R_2	25.391	19.249	19.448	21.196	12.616	15.091	16.456	23.532	25.287

Table 3.2: Components of the quadratic spline model for several breath cycles

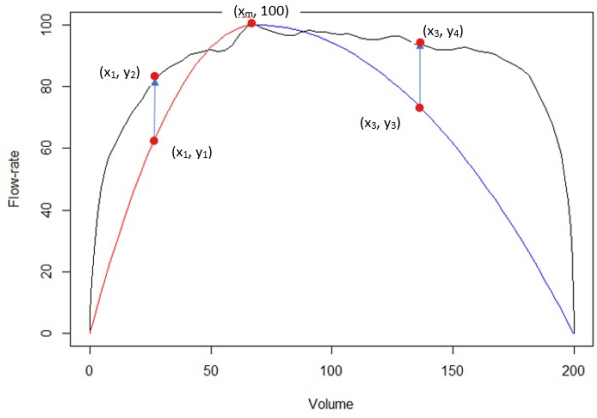


Figure 3.8: A quadratic spline model

3.2.3 Piecewise linear model

The third geometric model constructed in this study is a pieceswise linear model. To represent the model, a vector of twelve components is used. There are similarities and differences between the quadratic spline model and the piecewise linear model. Regarding the similarities, three identical marks appear in both models, which are the origin, maximum, and ending point. Thus, they share four identical components including the area a_u, ratios r_1 and r_3, and variable x_m, which are used to summarize the magnitude of the half-up cycles in both models.

The remaining eight elements of the vector representing the piecewise linear model are x_5, y_5, x_6, y_6, x_7, y_7, x_8, and y_8, defined as follows. x_5 and x_6 are the trisectors of the interval $[0, x_m]$, and y_5 and y_6 are the positive rescaled flow-rates at those two points. x_7, y_7, x_8, and y_8 are defined similarly on the interval $[x_m, 200]$.

Some calculation results for components of the piecewise linear model for a sample of 9 breath cycles are shown in Table 3.3. Figure 3.9 shows an example of piecewise linear model.

Cycle	1	2	3	4	5	6	7	8	9
a_u	118.421	95.891	102.516	128.721	134.847	90.893	76.869	46.454	113.084
r_1	4.692	4.878	4.951	4.749	4.970	6.158	6.015	6.085	4.916
r_3	31.107	35.327	32.678	27.387	25.078	31.321	36.282	56.925	30.077
x_m	68.489	78.312	71.445	62.884	55.164	76.806	59.341	108.827	76.281
x_5	12.334	14.266	15.794	12.045	9.618	14.053	10.135	13.967	12.041
y_5	64.349	64.353	67.841	62.814	58.202	66.070	70.695	55.883	58.372
x_6	38.052	43.943	40.095	36.532	29.014	42.909	33.847	56.775	41.580
y_6	88.298	93.210	91.673	84.850	88.758	93.928	91.627	90.118	90.876
x_7	118.926	128.949	124.314	116.884	112.021	120.120	118.672	150.671	128.690
y_7	95.101	91.230	90.812	95.317	86.163	90.714	93.023	88.000	94.297
x_8	170.133	174.218	172.727	167.945	160.524	165.163	167.043	183.292	176.051
y_8	89.114	73.406	74.158	82.096	82.416	93.928	76.277	68.589	73.483

Table 3.3: Components of the piecewise linear model for several breath cycles

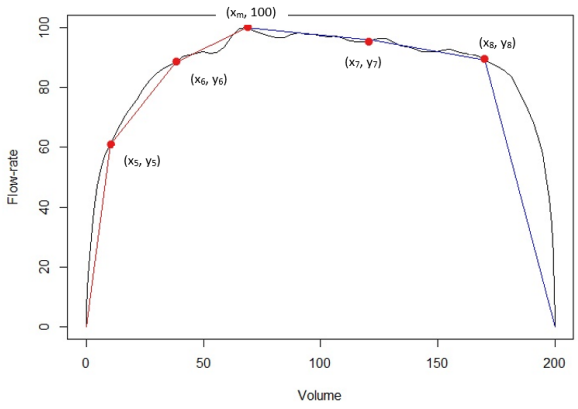


Figure 3.9: A piecewise linear model

3.3 Principal component analysis

Let X_0 be a $d \times n$ data matrix, to begin dimension reduction process, X_0 is normalized, which is described as

$$X = \frac{X_0 - \bar{X}}{sd(X_0)}$$

where $\hat{X}$ is the mean of the column of the matrix X_0, and $sd(X_0)$ is the standard deviation of the column of the matrix X_0.

Then, correlation matrix $\Sigma_{XX} = X^T X$ is constructed to describe linear dependence between variables in the matrix X. By diagonalizing the correlation matrix Σ_{XX}, we obtain loading W as eigenvectors, and variances Λ of new variables as eigenvalues, which is described as

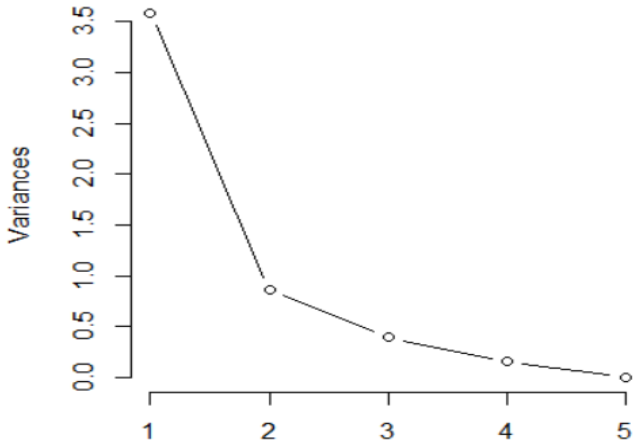
$$\Sigma_{XX} = W \Lambda W^T$$

where eigenvalues and eigenvectors are in the descending order of eigenvalues.

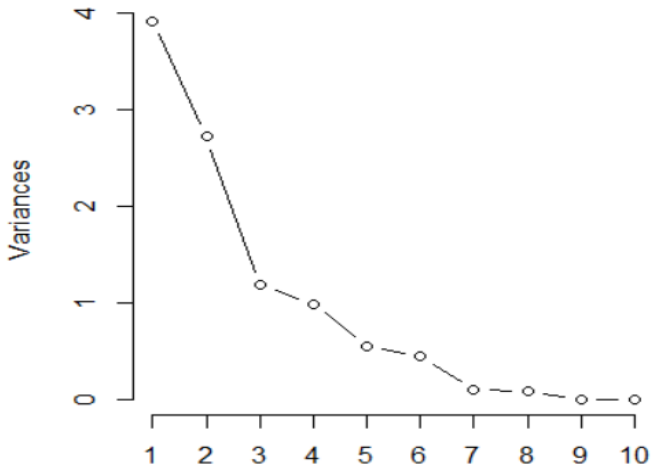
For each of the models described above, a model matrix X_0 is formed by using the vectors from all breath cycles as row vectors. The R function prcomp() is used to compute the singular value decomposition for these three model matrices.

By dropping some of the least important eigenvectors, PCA reduces the number of dimensions of the breath data. The number of dimensions can be determined by comparing the proportion of variation remaining in the models with the total variation of the dataset. Figure 3.10 shows three

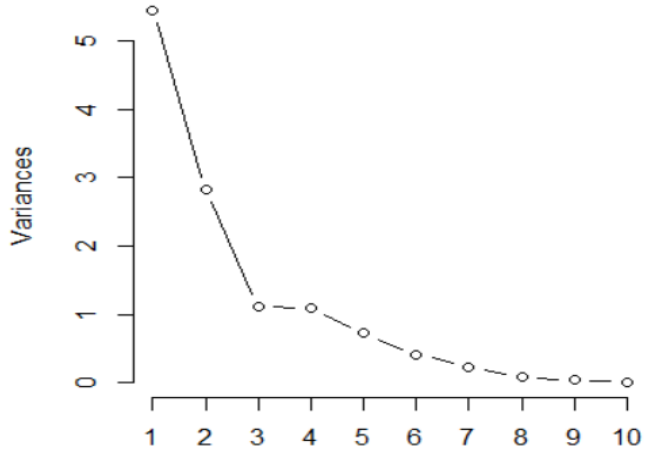
scree plots, which displays PCA results. A scree plot is a simple line segment plot, which shows the eigenvalues/variances of individual components in descending order [5].



(a) Scree plot from the simple oval model



(b) Scree plot from the quadratic spline model



(c) Scree plot from the piecewise linear model

Figure 3.10: PCA results from different models

To interpret the principal components (PCs) from each model, we look first at correlations between the breath data for each variable and each component, shown in Tables 3.4 to 3.6. From the correlations between variables and PCs, mathematical equations are constructed that can be used to calculate scores for healthy and testing animals.

According to the screen plots in Figure 3.10, the variances of PC1 and PC2 are significantly

higher than the other PCs. Tables 3.4 to 3.6 show the correlations between original variables and PCs from the models.

Variable	PC1	PC2
a_u	0.495	-0.033
r_1	-0.379	0.581
r_2	-0.400	0.493
r_3	-0.478	-0.445
r_4	-0.472	-0.469

Table 3.4: PCA results from the simple oval model

Two equations for computing $score_1$ and $score_2$ from the simple oval model are described as

$$score_1 = 0.495a_u - 0.379r_1 - 0.400r_2 - 0.478r_3 - 0.472r_4 \tag{3.11}$$

$$score_2 = -0.033a_u + 0.581r_1 + 0.493r_2 - 0.445r_3 - 0.469r_4 \tag{3.12}$$

Variable	PC1	PC2
a_u	0.149	-0.538
r_1	-0.059	0.429
r_3	-0.157	0.491
x_m	-0.473	-0.126
a_1	-0.433	0.024
b_1	0.461	0.020
a_2	0.352	0.296
b_2	-0.342	-0.302
R_1	0.088	-0.012
R_2	-0.277	0.298

Table 3.5: PCA results from the quadratic spline model

Two equations for computing $score_1$ and $score_2$ from the quadratic spline model are described as

$$\begin{aligned} score_1 = 0.149a_u - 0.059r_1 - 0.157r_3 - 0.473x_m - 0.433a_1 + 0.461b_1 + 0.352a_2 \\ - 0.342b_2 + 0.088R_1 - 0.277R_2 \end{aligned} \tag{3.13}$$

$$score_2 = -0.538a_u + 0.429r_1 + 0.491r_3 - 0.126x_m + 0.024a_1 + 0.020b_1 + 0.296a_2 - 0.302b_2 - 0.012R_1 + 0.298R_2 \tag{3.14}$$

Variable	PC1	PC2
a_u	0.085	-0.550
r_1	0.004	0.438
r_3	-0.098	0.510
x_m	-0.419	-0.057
x_5	-0.404	-0.153
y_5	-0.233	-0.333
x_6	-0.417	-0.095
y_6	0.031	-0.085
x_7	-0.423	-0.013
y_7	-0.173	0.142
x_8	-0.413	0.093
y_8	0.188	-0.252

Table 3.6: PCA results from the piecewise linear model

Two equations for computing $score_1$ and $score_2$ from the piecewise linear model are described as

$$score_1 = 0.085a_u + 0.004r_1 - 0.098r_3 - 0.419x_m - 0.404x_5 - 0.233y_5 - 0.417x_6 + 0.031y_6 - 0.423x_7 - 0.173y_7 - 0.413x_8 + 0.188y_8 \tag{3.15}$$

$$score_2 = -0.550a_u + 0.438r_1 + 0.510r_3 - 0.057x_m - 0.153x_5 - 0.333y_5 - 0.095x_6 - 0.085y_6 - 0.013x_7 + 0.142y_7 + 0.093x_8 - 0.252y_8 \tag{3.16}$$

3.4 Distributions of scores for healthy animals

In this section, to determine scores characterizing the breath cycle of healthy animals, a set of 92 breath cycles taken from five healthy animals are combined into a training set referred to below as **animal H**. Then, we construct boxplots and, quantile-quantile (Q-Q) plots, and bagplots to describe empirical distributions of scores from the breath data of animal H. From the boxplot, we

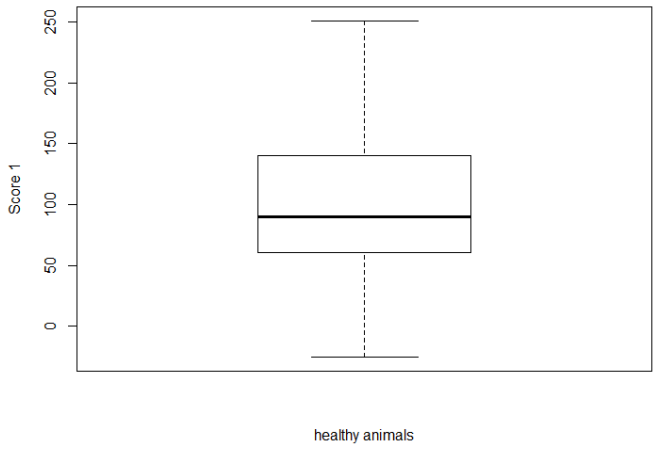
can visualize the level of symmetry or skewness of distribution of each score from the breath data of healthy animals [10]. Abundant applications of boxplot are discussed in the article of Williamson, in which his main interest is how boxplot can interpret a dataset from complex table [25]. Besides that, Q-Q plot displays quantiles of one dataset against quantiles of a second dataset [9]. The main purpose of Q-Q plot is to determine whether two datasets have the same distribution. The interpretation process of Q-Q plot is described in detail by Wang, Steele, and Zhang [24]. In our study, the first dataset is scores from the breath data of healthy animals with unknown distribution, and the second dataset has the standard normal distribution. As a result, if the breath data of healthy animals are from normal distribution, data points will position close to Q-Q line that is a straight diagonal line. In addition, bivariate boxplot displays two-dimensional distributions of multivariate dataset [18]. From the bivariate boxplot, we can visualize the spread and outlier of the breath data. We also use the Shapiro-Wilk test and Anderson-Darling test to test for normality [20] [2].

Figures 3.11 to 3.13 show distributions of each score from breath data of healthy animals. Figures 3.14 show two-dimensional distributions of scores from breath data of healthy animals from the three models. In addition, Table 3.7 shows the normality test results.

Model	score	Shapiro-Wilk test	Anderson-Darling test
Simple oval	$score_1$	0.082	0.029
	$score_2$	0.616	0.603
Quadratic spline	$score_1$	0.836	0.934
	$score_2$	0.286	0.314
Piecewise linear	$score_1$	0.100	0.264
	$score_2$	0.970	0.950

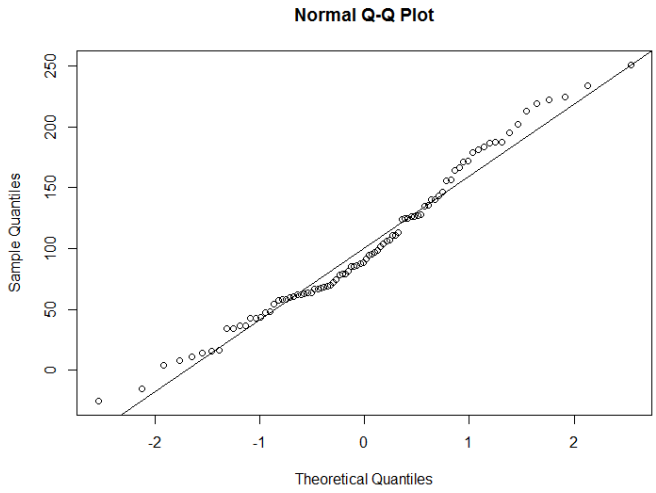
Table 3.7: Normality test results from different models

According to Figures 3.11 to 3.14, neither plots reveal unusual qualities, such as outliers or gaps. The boxplots show that the distributions of scores are slightly skewed. Even though there are a few points wriggle about the Q-Q lines in the Q-Q plots, the majority lie quite close to the Q-Q lines. Besides that, from the bivariate boxplots, their means are located approximately at the centers of the boxplots, and the spreads of the data are small as the majority lie relatively close to the means. In addition, according to the normality test results from Shapiro-Wilk and Anderson-

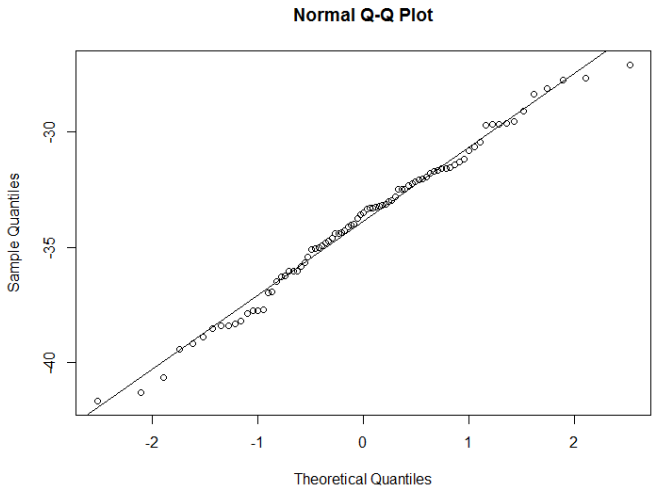


(a) $Score_1$

(b) $Score_2$

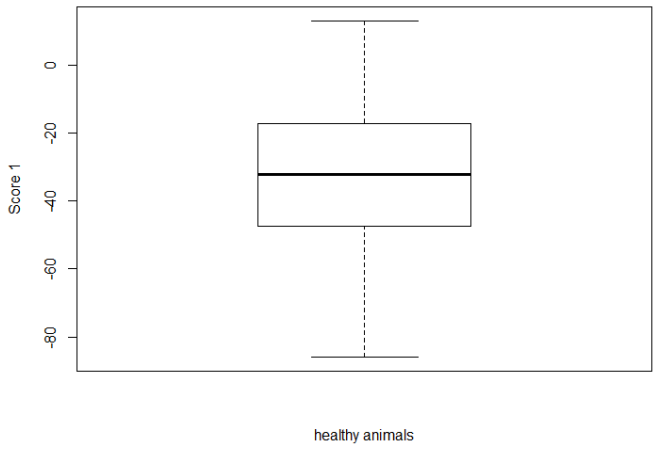


(c) $Score_1$



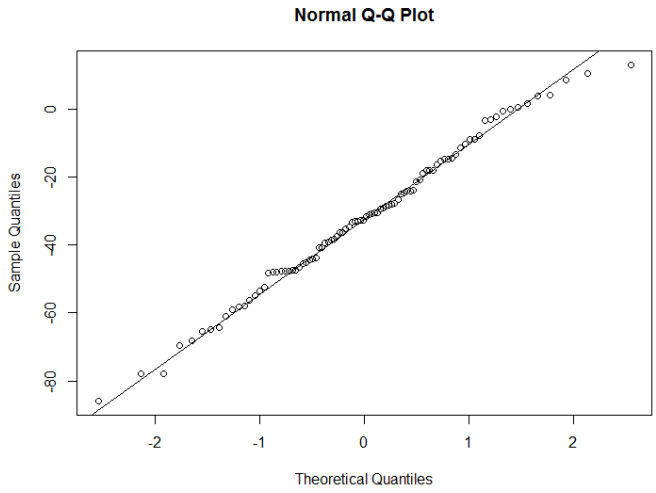
(d) $Score_2$

Figure 3.11: Boxplots and Q-Q plots of the first two scores from the simple oval model for healthy animals

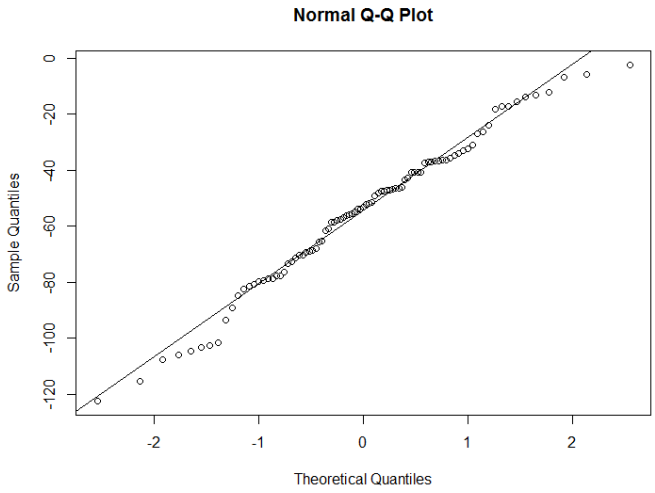


(a) $Score_1$

(b) $Score_2$

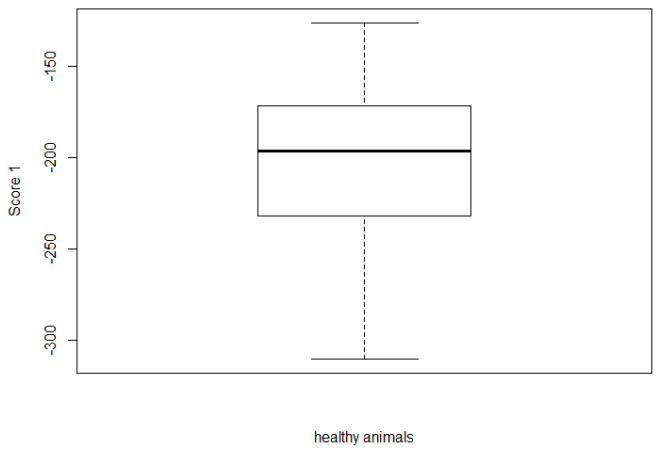


(c) $Score_1$

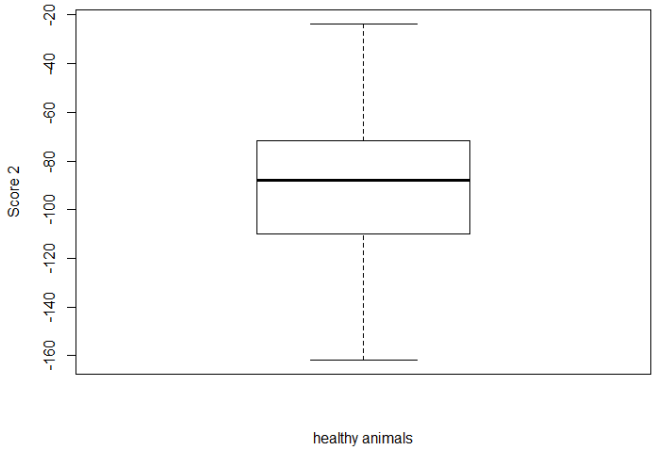


(d) $Score_2$

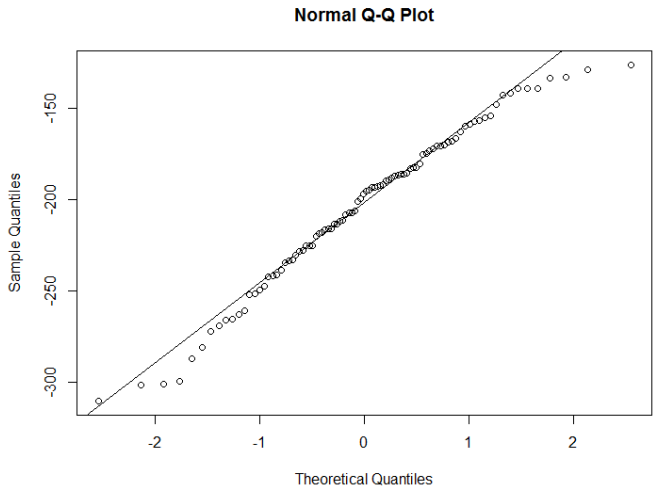
Figure 3.12: Boxplots and Q-Q plots of the first two scores from the quadratic spline model for healthy animals



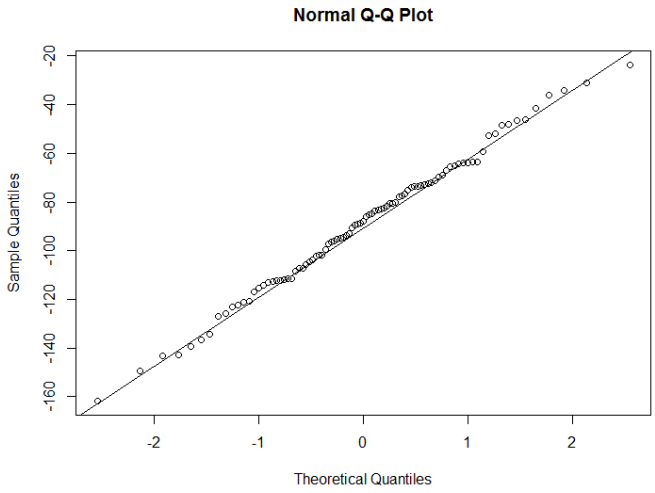
(a) $Score_1$



(b) $Score_2$

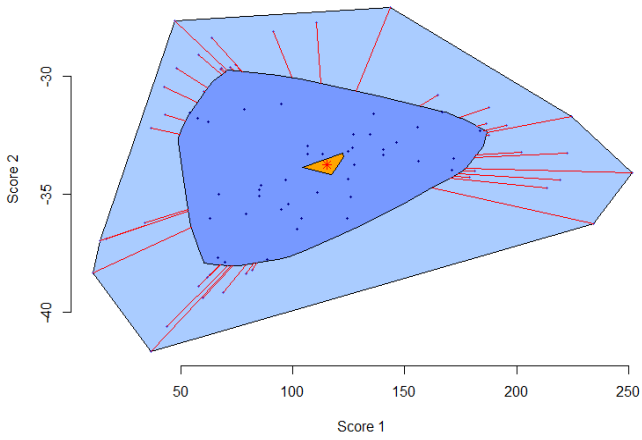


(c) $Score_1$

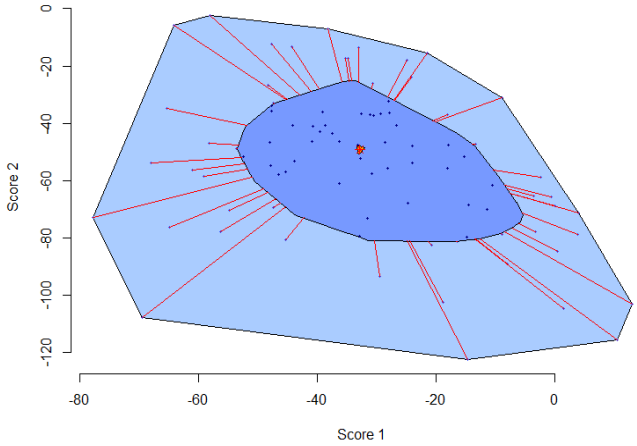


(d) $Score_2$

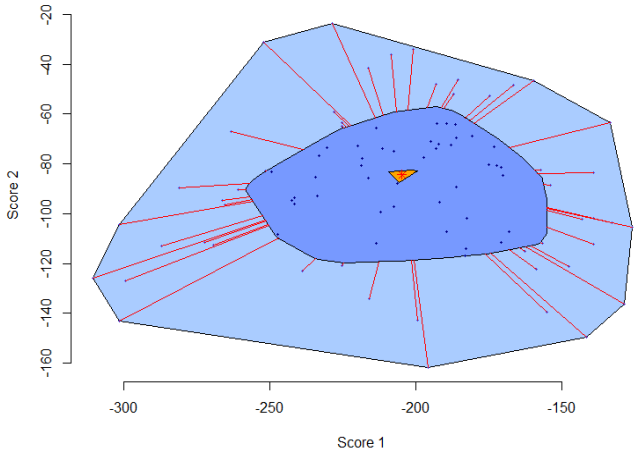
Figure 3.13: Boxplots and Q-Q plots of the first two scores from the piecewise linear model for healthy animals



(a) The simple oval model



(b) The quadratic spline model



(c) The piecewise linear model

Figure 3.14: Bivariate boxplots of the first two scores from different models for healthy animals

Darling tests, all p-values are greater than the 1% significant level. These results/findings indicate that the distributions of scores from the breath data of healthy animals are approximately normal. In the future, we will use two-dimensional normality tests on scores of healthy animals for validating the bivariate normality.

3.5 Criteria for disease diagnosis

In this section, based on the PCs and scores collected from PCA, we define criteria for disease diagnosis. Because the breath data of different animals have different numbers of breath cycles measured, we define criteria for disease diagnosis according to the number of breath cycles. Firstly, if a dataset from a given animal contains only one breath cycle and its scores are beyond the normal ranges of the healthy population, there are some differences between healthy and testing animals. Secondly, if a dataset from a given animal contains a few breath cycles and its average scores are beyond the normal ranges of the healthy population, there are some differences between healthy and testing animals. The normal ranges of the healthy population are defined by 3 standard deviations from the mean criterion or 1.5 interquartile range criterion in a modified boxplot. Finally, if a dataset from a testing animal contains many cycles, two-sample t-test and Hotelling's t-squared test are used to test the difference between healthy and testing animals [22] [7]. While the two-sample t-test is used to compare the two population means of one score from healthy and testing animals, the Hotelling's t-squared test is used to test the mean vector of the two sets of score values from healthy and testing animals. The equation for computing diseased index for the two-sample t-test is defined as

$$\text{Diseased index T} = \frac{\overline{\mathbf{x}} - \overline{\mathbf{y}}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.17}$$

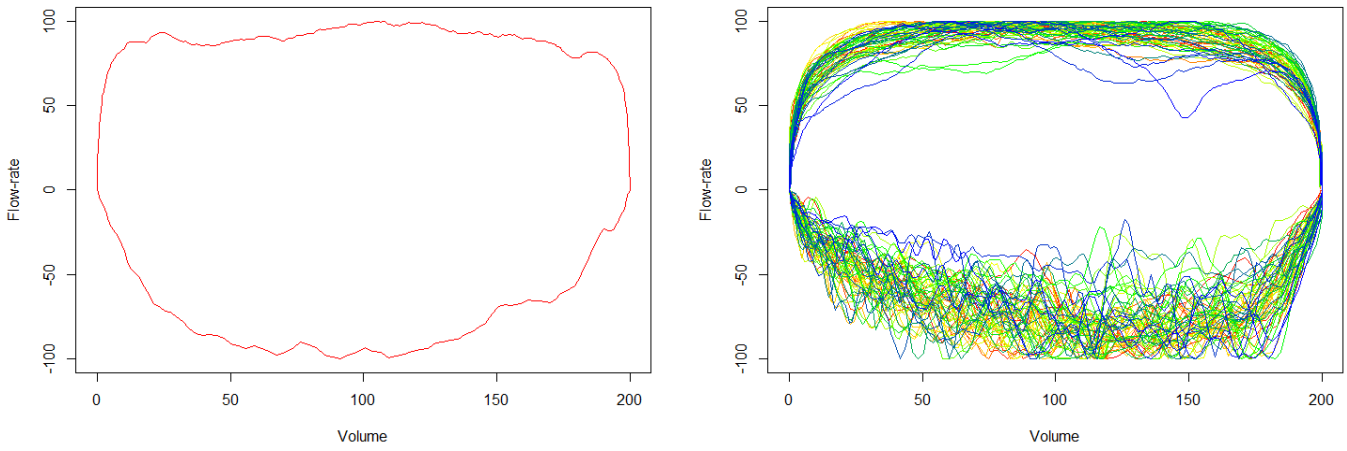
where $\overline{\mathbf{x}}$ is the average score of a testing animal, $\overline{\mathbf{y}}$ is the average score of healthy animals, n_1 is the number of breath cycles from a testing dataset, n_2 is the number of breath cycles from healthy dataset, s_1 is the standard deviation of scores of a testing animal, and s_2 is the standard deviation of scores of healthy animals.

The equation for computing diseased index for the Hotelling's t-squared test is defined as

$$\text{Diseased index } T^2 = (\overline{\mathbf{x}} - \overline{\mathbf{y}})^T (\hat{\Sigma}(\frac{1}{n_1} + \frac{1}{n_2}))^{-1} (\overline{\mathbf{x}} - \overline{\mathbf{y}}) \tag{3.18}$$

where $\overline{\mathbf{x}} = \frac{1}{n_x} \sum_{i=1}^{n_x} \mathbf{x}_i$ is the average score vector of a testing animal, $\overline{\mathbf{y}} = \frac{1}{n_y} \sum_{i=1}^{n_y} \mathbf{y}_i$ is the average score vector of healthy animals, and $\hat{\Sigma} = \frac{n_x \hat{\Sigma}_{\mathbf{x}} + n_y \hat{\Sigma}_{\mathbf{y}}}{n_x + n_y - 2}$ is the covariance matric where $\hat{\Sigma}_{\mathbf{x}} = \frac{1}{n_x - 1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$ is the covariance matrix of a testing animal and $\hat{\Sigma}_{\mathbf{y}} = \frac{1}{n_y - 1} \sum_{i=1}^{n} (\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})'$ is the covariance matrix of healthy animals. Under the null hypothesis, the statistics T^2 is related to the F-distribution with p and $n - p$ degrees of freedom, where $p = 2$ for comparing bivariate scores.

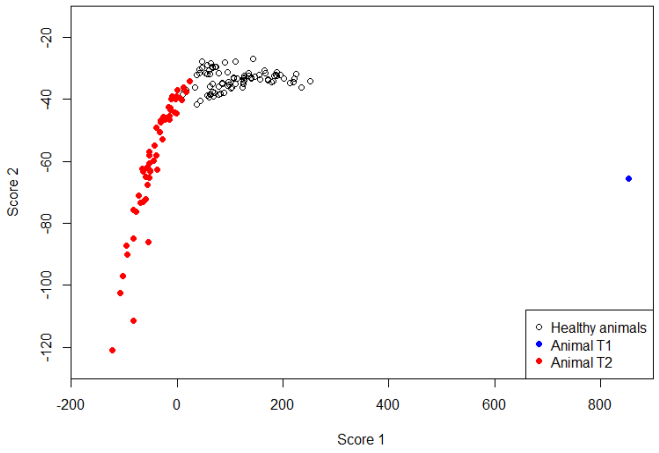
In this chapter, based on the scores defined in equations 3.12 to 3.16, we compute scores for testing animals. Then we test the claim that the mean for a testing animal is the same as, or different from, the mean for the healthy population. Results from the test will be used to detect the health condition of two testing animals which we will refer to as **animal T1** and **animal T2**. T1 has one breath cycle measured, while T2 has 59 breath cycles measured. In addition, scatter plots are used to compare the two-dimensional measurements from healthy animals with testing animals, as shown in Figure 4.16. Figure 4.15 shows breath cycles of T1 and T2 after curve registration. We also use scatter plot to display the distribution of scores from healthy and testing animals, which is shown in figure 4.16.



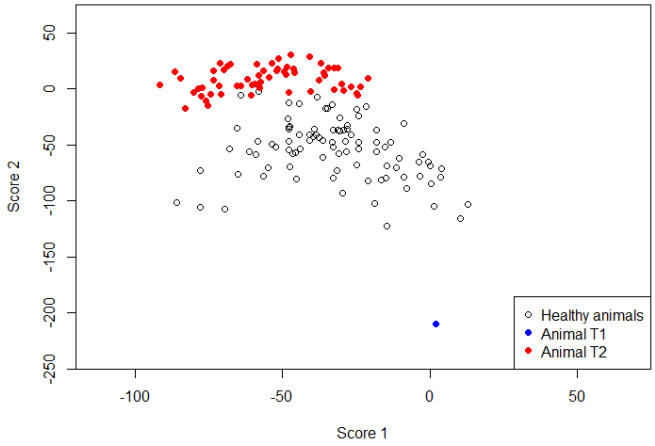
(a) Animal T1 (b) Animal T2

Figure 4.15: Breath cycles of testing animals after curve registration

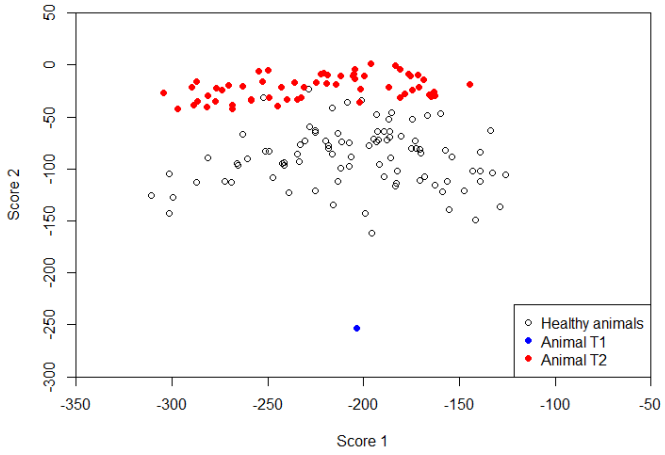
For the disease diagnosis for animal T1, since it has one breath cycle measured, we plot the single point for animal T1 against all the points for animal H, as shown in Figure 4.17. According to Figures 4.16 and 4.17, the scores of animal T1 appear to be outliers in the plots. That indicates some differences between healthy animals and animal T1.



(a) The simple oval model



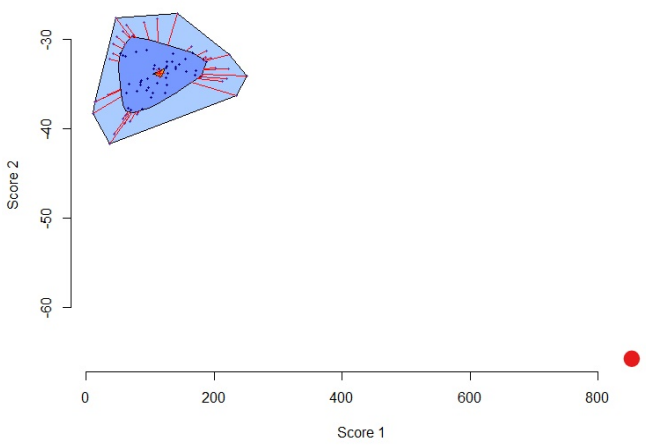
(b) The quadratic spline model



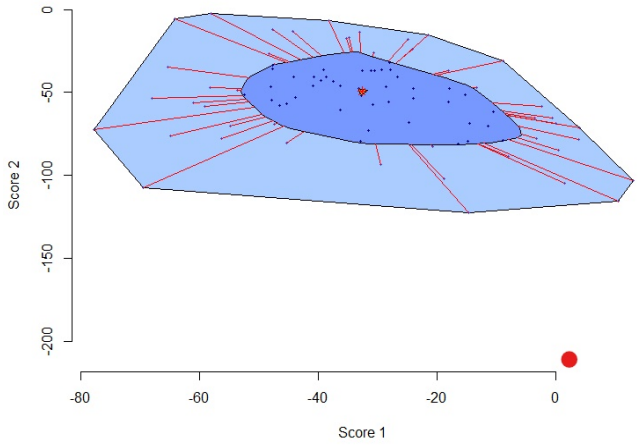
(c) The piecewise linear model

Figure 4.16: Scatter plots of the first two scores from different models for healthy and testing animals

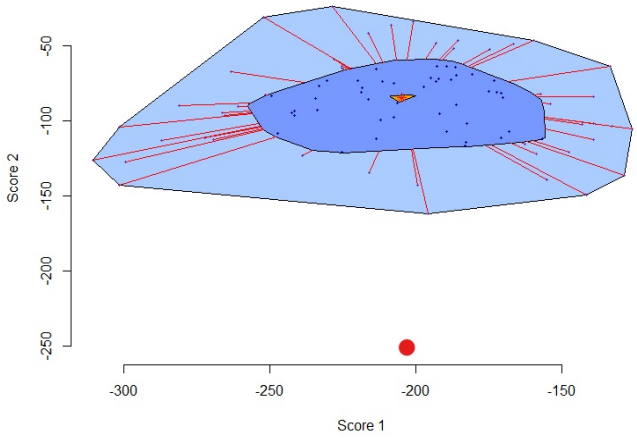
For the disease diagnosis for animal T2, since it has 59 breath cycles measured, we use two-sample t-test and Hotelling's t-squared test to test the differences between healthy animals and testing animal T2. Two-sample t-test is used to compare the two population means of one score from healthy animals and animal T2. The null hypothesis H_0 states that the mean for animal T2 is the same as the mean for the healthy population. The alternative hypothesis H_1 states that the mean for animal T2 is different from the mean for the healthy population. Table 4.8 shows results



(a) The simple oval model



(b) The quadratic spline model



(c) The piecewise linear model

Figure 4.17: Bivariate boxplots of the first two scores from different models for animal H and animal T1

from the two-sample t-test.

Hotelling's t-squared test is used to test the mean vector of the two sets of score values from healthy animals and animal T2. The null hypothesis H_0 states that the mean vector for testing animal is the same as the mean vector for the healthy population. The alternative hypothesis H_1 states that the mean vector for animal T2 is different from the mean vector for the healthy population. Table 4.9 shows results from the Hotelling's t-squared test.

Model	score	p-value
Simple oval	$score_1$	$<$ 2.2e-16
	$score_2$	1.262e-13
Quadratic spline	$score_1$	2.449e-11
	$score_2$	$<$ 2.2e-16
Piecewise linear	$score_1$	0.004
	$score_2$	$<$ 2.2e-16

Table 4.8: Two-sample t-test results from different models

Model	p-value
Simple oval	$<$2.2e-16
Quadratic spline	$<$2.2e-16
Piecewise linear	$<$2.2e-16

Table 4.9: Hotelling's t-squared test results from different models

As can be seen from Tables 4.8 and 4.9, all the p-values are less than the 1% significant level, so we reject the null hypothesis H_0 [22] [7]. In addition, according to Figure 4.16, the distributions of scores for animal T2 are separated from the distributions of scores for healthy animals. These findings/results indicate some differences between healthy animals and animal T2.

CHAPTER V: CONCLUSION AND FUTURE WORK

We propose some statistical methods for disease diagnosis on dolphin by analyzing the breath data. In the analysis, a total of 92 breath cycles of five healthy dolphins have been analyzed to build the baseline of healthy animals.

The ratios/area are used to describe the magnitude of the breath curves. Three models are constructed to describe the geometric shape of the breath curves. Then PCA are applied to extract important features from the magnitude/shape descriptors. Criteria for disease diagnosis are defined using boxplots and bivariate boxplots to display the distributions of scores for healthy and testing animals, the two-sample t-test to compare the two population means of one score from healthy and testing animals, and the Hotelling's t-squared test to test the mean vector of the two sets of score values from healthy and testing animals. The proposed methods were applied to detect the health condition of two testing animals. The results/findings indicate that there are some differences between healthy and testing animals.

For future work, we plan to focus on the analysis of the half-down cycles, where a large amount of variation provides a rich source of information. In addition, we propose to continue collecting and analyzing more breath data to check the precision of our method. Finally, more techniques should be used to explore the potential information covered in the breath data of marine mammals.

REFERENCES

[1] Hervé Abdi and Lynne J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 7 2010.

[2] Anderson, T.W, Darling, and D. A. Asymtotic theory of certain goodness-of-fit. *Annals Mathematical Statistocs*, 23:193–212, 1952.

[3] Eric Booth, Jeff Mount, and Joshua H. Viers. Hydrologic variability of the cosumnes river floodplain. *San Francisco Estuary and Watershed Science*, 4(2), 2006.

[4] Richard Dehn and David P. Asprey. *Essential clinical procedures*. Elsevier Saunders, Philadelphia, PA, 3 edition, 2013.

[5] Valerie J. Easton and John McColl. Scree plot. The Statistics Glossary website.

[6] Steven M. Holland. Principal components analysis (pca). Technical report, Department of Geology, University of Georgia, Athens, GA, 5 2008.

[7] H Hotelling. The generalization of student's ratio. *Ann Math Stat*, 3(2):360–378, 1931.

[8] George B. Thomas Jr. and Ross L. Finney. *Calculus and analytic geometry*. Addison-Wesley, 9 edition, 1996.

[9] Henry C. Thode Jr. *Testing for normality*, volume 164. Marcel Dekker, Newyork, 2002.

[10] Robert Mc.Gill, John W. Tukey, and Wayne A. Larsen. The title of the work. *The name of the journal*, 32(1):12–16, 2 1978.

[11] Francesca Montani, Matteo Jacopo Marzi, Fabio Dezi, Elisa Dama, Rose Mary Carletti, Giuseppina Bonizzi, Raffaella Bertolotti, Massimo Bellomi, Cristiano Rampinelli, Patrick Maisonneuve, Lorenzo Spaggiari, Giulia Veronesi, Francesco Nicassio, Pier Paolo Di Fiore, and Fabrizio Bianchi. mir-test: A blood test for lung cancer early detection. *J Natl Cancer Inst*, 107(6), 3 2015.

[12] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[13] J. O. Ramsay. When the data are functions. *Psychometrika*, 47(4):379–396, 12 1982.

[14] J. O. Ramsay and Xiaochun Li. Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):351–363, 1998.

[15] James O. Ramsay and Bernald W. Silverman. *Applied functional data analysis: methods and case studies*, volume 77. Springer, New Yorl, 2002.

[16] C. Radhakrishna Rao. Some statistical methods for comparison of growth curves. *Biometrics*, 14(1):1–17, 3 1958.

[17] Markus Ringnér. What is principal component analysis? *Nature Biotechnology*, 26(3):303–304, 3 2008.

[18] Peter J. Rousseeuw, I. Ruts, and J.W Tukey. The bagplot: A bivariate. *The American Statistician*, 53(4):382–387, 1999.

[19] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 2 1978.

[20] Shapiro, S.S, and MB Wilk. An analysis of variance test for normality. *Biometrika*, 52(3):591–593, 1965.

[21] B. W. Silverman. Incorporating parametric effects into functional principal components analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(4):673–689, 1995.

[22] George W. Snedecor and William G. Cochran. Statistical methods. *Iowa State University Press*, 1989.

[23] T.B. Waltzek, G. Cortes-Hinojosa, J.F. Wellehan Jr, and G.C. Gray. Marine mammal zoonoses: A review of disease manifestations. *Zoonoses and Public Health*, 59:521–535, 2012.

[24] Yunsi Wang, Tyler Steele, and Eva Zhang. Qq plot. 2016.

[25] D. F. Williamson, R. A. Parker, and J. S. Kendrick. The box plot: a simple visual method to interpret data. *Ann Intern Med*, 110(11):916–921, 6 1989.

APPENDIX A: R CODES

```
#Code1: plotting original breath cycles from raw data
file_name = c() #list of file names
num_data = #the file number
dat = read.csv(paste(file_name[num_data],".csv",sep=""))
plot(dat[start.cycle:end.cycle,2], dat[start.cycle:end.cycle,1],type = "l")

#Code2: decomposing time series into individual breath cycles
attach(dat)
Min.interval.bw.two.cycles = 0.5
diff.Flowrate = c(0, diff(Flowrate))
max.Flowrate = max(Flowrate)
min.Flowrate=min(Flowrate)
max_min = max.Flowrate - min.Flowrate
#To select thresholds to determine if a value is in a breathe cycle
#or not by the size of max breathe cycle
k1 = min(max(max_min/80,0.09), 0.5)
k2 = min(max(max_min/80,0.09),0.25)
k3= min(max(max_min/40,0.2),1.25)
k4.max.min= max(max_min/20, 2)
#To identify possible values in breath cycles by its absolute values and the change
#If the difference = 0, then it is very likely not in a breathe cycle
#It the values in a consequence of several are very small, it is not in a cycle
T = length(Flowrate)
index_noneZero= ((abs(Flowrate)>k1) & (abs(diff.Flowrate)>k2) | (abs(Flowrate)>k3))
```

```
index_noneZero.corrected = index_noneZero
#To define the i-th plausible cycle (possible false cycles )
#If there is a transition from 1 to 0 (at least a consequence of 10 0s),
#the next cycle to appear soon and icycle_value = icycle_value + 1
i_cycle = rep(0, T)
i_cycle_value = 1
max.T.Cyl = 0
T.cyl =0
DistBwCycles= 0
start.cycle=list()
end.cycle=list()
for( i in 1: (T-5))
{
    if (index_noneZero.corrected[i] == TRUE)
    {
        if(T.cyl == 0)  { start.cycle[  i_cycle_value ] = i}
        i_cycle[i] = i_cycle_value
        T.cyl = T.cyl + 1
        if( (sum(index_noneZero.corrected[(i+1):(i+10)]) == FALSE) )
        {
            end.cycle[  i_cycle_value ] = i
            i_cycle_value = i_cycle_value + 1
            max.T.Cyl = max(max.T.Cyl,  T.cyl )
            T.cyl  = 0
        }
    }
}
```

```
#Code3: removing incomplete cycles
file_name = c() #list of file names
num_data = #the file number
dat = read.csv(paste(file_name[num_data],".csv",sep=""))
#enter estimated starting points of individual cycles,
#here the points are gradually increased until flat segments are eliminated
start.cycle = ()
#enter estimated ending points of individual cycles,
#here the points are gradually decreased until flat segments are eliminated
end.cycle = ()
#check visually by plots
plot(dat[start.cycle:end.cycle,2], dat[start.cycle:end.cycle,1],type = "l")

#Code4: moving averages
dat = read.csv(paste(file,"_dat.csv",sep=""))
max_cycle = dat$cycle[nrow(dat)]
new_dat = dat[c(),]
for(cycle in 1:max_cycle){
    temp = dat[dat$cycle==cycle,]
    temp1 = temp[temp$side == "up",]
    temp1 = temp1[order(temp1$TidalVolume),]
    temp2 = temp[temp$side == "down",]
    temp2 = temp2[order(temp2$TidalVolume),]
    n = nrow(temp1)
    temp1$Flowrate[1:(n-1)] = (temp1$Flowrate[1:(n-1)] + temp1$Flowrate[2:n])/2
    temp1$TidalVolume[1:(n-1)] = (temp1$TidalVolume[1:(n-1)] +
```

```
temp1$TidalVolume[2:n])/2
    new_dat[(nrow(new_dat)+1):(nrow(new_dat)+n-1),] = temp1[1:(n-1),]
    n = nrow(temp2)
    temp2$Flowrate[1:(n-2)] = (temp2$Flowrate[1:(n-2)] + temp2$Flowrate[2:(n-1)]
+ temp2$Flowrate[3:n])/3
    temp2$TidalVolume[1:(n-2)] = (temp2$TidalVolume[1:(n-2)]
+ temp2$TidalVolume[2:(n-1)] + temp2$TidalVolume[3:n])/3
    new_dat[(nrow(new_dat)+1):(nrow(new_dat)+n-2),] = temp2[1:(n-2),]
}
write.csv(new_dat,paste(file,"_dat2.csv",sep=""),row.names = FALSE)

#Code5: rescaling breath cycles
for(file in list_file){
    if(new_name){
        dat = read.csv(paste(file,"_dat2.csv",sep=""))
    }else{
        dat = read.csv(paste(file,"_dat.csv",sep=""))
    }
    n = dat$cycle[nrow(dat)]
    for (i in 1:n){
        temp = dat[dat$cycle==i,]
        scale_dat[current,"Name"] = file
        scale_dat[current,"cycle"] = i
        scale_dat[current,"up_scale"] = 100/max(temp[temp$side=="up","Flowrate"])
        scale_dat[current,"down_scale"] = -100/min(temp[temp$side=="down","Flowrate"])
        scale_dat[current,"x_up_scale"] = 200/max(temp[temp$side=="up","TidalVolume"])
        scale_dat[current,"x_down_scale"] = 200/max(temp[temp$side=="down","TidalVolume"
```

```
        current = current+1
        temp[temp$side=="up","Flowrate"] = temp[temp$side=="up","Flowrate"]
*100/max(temp[temp$side=="up","Flowrate"])
        temp[temp$side=="down","Flowrate"] = -temp[temp$side=="down","Flowrate"]
*100/min(temp[temp$side=="down","Flowrate"])
        temp[temp$side=="up","TidalVolume"] = temp[temp$side=="up","TidalVolume"]
*200/max(temp[temp$side=="up","TidalVolume"])
        temp[temp$side=="down","TidalVolume"] = temp[temp$side=="down","TidalVolume"]
*200/max(temp[temp$side=="down","TidalVolume"])
        dat[dat$cycle==i,] = temp
    }
    if(new_name){
        write.csv(dat,paste(file,'_std2.csv',sep=""),row.names = FALSE)
    }else{
        write.csv(dat,paste(file,'_std.csv',sep=""),row.names = FALSE)
    }
}
```