University of New Orleans
ScholarWorks@UNO

University of New Orleans Theses and Dissertations | Dissertations and Theses

8-5-2010

The Distribution of Cotton Fiber Length

Rachid Belmasrour
University of New Orleans

Follow this and additional works at: https://scholarworks.uno.edu/td

Recommended Citation

Belmasrour, Rachid, "The Distribution of Cotton Fiber Length" (2010). *University of New Orleans Theses and Dissertations*. 1216.
https://scholarworks.uno.edu/td/1216

This Dissertation is protected by copyright and/or related rights. It has been brought to you by ScholarWorks@UNO with permission from the rights-holder(s). You are free to use this Dissertation in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you need to obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/or on the work itself.

This Dissertation has been accepted for inclusion in University of New Orleans Theses and Dissertations by an authorized administrator of ScholarWorks@UNO. For more information, please contact scholarworks@uno.edu.

The Distribution of Cotton Fiber Length

A Dissertation

Submitted to the Graduate Faculty of the
University of New Orleans
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Engineering and Applied Science

by

Rachid Belmasrour
B.A. University of Hassan II, 1998
M.S., University of Versailles Saint Quantin, 2001
M.S., University of New Orleans, 2007

August, 2010

Copyright 2010,
RACHID BELMASROUR

Dedication

This thesis is dedicated to my father, who taught me that the best kind of knowledge to have is that which is learned for its own sake. It is also dedicated to my mother, who taught me that even the largest task can be accomplished if it is done one step at a time. I also dedicate it to my dear wife, who remains willing to engage with the struggle, and ensuing discomfort. Your unconditional love and support has overfilled my heart with happiness and joy. You understand and know me inside and out.

Acknowledgments

I would like to express my sincere appreciation to my advisor Dr. Li, Linxiong for suggesting this research project, his advice during the course of my graduate studies, and his inspiration and guidance which made it possible for me to achieve this stage. My special gratitude goes to Dr. Xiaoliang Cui for his support and valuable discussion. I also thank Dr. Tumulesh Solanky, Dr. Jairo Santanilla, and Dr. Vesselin Jilkov for their time and patience in reading and reviewing my dissertation.

I am grateful to the Department of Mathematics and United States Department of Agriculture for the support of this research.

I would like to thank my dear wife, Nancy and my family for their love, understanding and support.

Contents

Abstract

By testing a fiber beard, certain cotton fiber length parameters can be obtained rapidly. This is the method used by the High Volume Instrument (HVI). This study is aimed to explore the approaches and obtain the inference of length distributions of HVI beard samples in order to develop new methods that can help us find the distribution of original fiber lengths and further improve HVI length measurements. At first, the mathematical functions were searched for describing three different types of length distributions related to the beard method as used in HVI: cotton fiber lengths of the original fiber population before picked by the HVI Fibrosampler, fiber lengths picked by HVI Fibrosampler, and fiber beard's projecting portion that is actually scanned by HVI. Eight sets of cotton samples with a wide range of fiber lengths are selected and tested on the Advanced Fiber Information System (AFIS). The measured single fiber length data is used for finding the underlying theoretical length distributions, and thus can be considered as the population distributions of the cotton samples. In addition, fiber length distributions by number and by weight are discussed separately. In both cases a mixture of two Weibull distributions shows a good fit to their fiber length data. To confirm the findings, Kolmogorov-Smirnov goodness-of-fit tests were conducted. Furthermore, various length parameters such as Mean Length (ML) and Upper Half Mean Length (UHML) are compared between the original distribution from the experimental data and the fitted distributions. The results of these obtained fiber length distributions are discussed by using Partial Least Squares (PLS) regression, where the distribution of the original fiber length from the distribution of the projected one is estimated.

Finally, reducing the number of parameters in a regression can enhance the estimation of parameters. To this end we introduced a new distribution with only three parameters to describe the distribution of fiber lengths by weight.

KEYWORDS: Fiber Beard, Komogorov-Simirnov goodness-of-fit test, Mixture of Weibull Distributions, Partial Least Squares.

List of Figures

List of Tables

1

Chapter 1

1.1 Introduction

Fiber length is considered the most important property of cotton in marketing and yarn processing. In the past decades, cotton industry and researchers have been trying to develop efficient methods to measure the length parameters of cotton fiber. These parameters include Mean Length (ML), Upper Half Means Length (UHML), Short Fiber Content (SFC), Uniformity Index (UI), etc. Measuring a fiber beard by using the High Volume Instrument (HVI) instead of individual fibers provides a quick solution for those fiber length parameters, (Suh and Sasser 1996). In HVI testing, the specimen fibers are picked up by the needles on the comb/clamp through holes of the HVI Fibrosampler. The specimen fibers are in the form of a tapered beard. The curve obtained from the measurement, so-called Fibrogram, describes the relationship of length and density of this tapered beard.

The original theory of the Fibrogram as developed by Hertel (1936, 1940) has served as the basis of subsequent cotton length measurement methods based on fiber beards. Following Hertel's pioneer work, various developments have been made. (Krowicki et al. 1996) generated distributions from cotton fibers Fibrogram. The generated fiber lengths were presented as graphical bar charts in discrete form, and not as mathematical functions.

Early investigations of (Prier and Sasser 1971) discussed three different theoretical fiber length distribution density functions: a uniform density and two triangular densities. They claimed that one of the triangular densities could be used to describe short fiber lengths, the

other triangular density could be used for long fiber lengths, and the uniform density could be used for middle fiber lengths. They further stated that a mixture of these three densities can closely match any set of measured data. However, they did not provide a method to mix those densities. Instead, they concluded that it was not feasible to obtain an explicit expression for the probability density function of the whole fiber length.

Furthermore, (Zeidman et al. 1991) discussed the range and shape of experimental length distributions and their relationships to length parameters. They found that one single parameter could not sufficiently characterize the entire fiber length distribution and concluded that more statistical measures are needed for distribution location, dispersion, and shape.

In an attempt to describe cotton fiber length distribution, (Krifa) 2006 studied the modality of fiber length distribution and relationships between modality and other cotton properties such as maturity and strength. In his later reports (2007 & 2008) mixed Weibull distribution was utilized to describe cotton fiber length and parameters. Krifa's main focus was on the changing of modality during processes.

Other efforts focused on the estimation of statistics from Fibrograms and the difference between length distribution by number and by weight. (Cui, Calamari and Suh 1998) showed that when comparing two cottons, this difference may give different rank orders. Sampling method was discussed regarding its impact on how to explain the relationship between the original fiber length distribution and the Fibrogram obtained from HVI.

1.2 **Scope of research**

This study is aimed at exploring several approaches to obtain inference of length distributions of a sample beard as used by HVI and thus investigating its relationship with the actual fiber length distribution that can be obtained from other test methods and devices such as AFIS. This can provide information that will help to develop new length parameters

and as a result better suit the needs of the industry, which in turn expand the utilization of HVI results. In addition, it will help in the understanding of the different measurement results between HVI and AFIS, which have been reported by earlier research (Cui 1997). To achieve such objective, mathematical functions that describe the underlying population distributions of the fiber lengths related to HVI measurements must be established. In other words, if the distribution function is known, then all the length parameters can be calculated.

Mixed Weibull distribution was used to describe cotton fiber length. Apparently, this distribution requires quite a number of parameters, and these parameters could be in nonlinear forms which will make the estimation of the distribution and matching extremely difficult. This study focuses on finding and validating the distribution functions dealing with three different types of fiber length distributions that are related to HVI measurements.

In practice, the distribution of projected fiber lengths is the only available distribution; on the other hand the original fiber length is unknown and unavailable. To better understand the quality of cotton, one needs to know the length distribution of the original fiber and thus would like to obtain the distribution of the original fiber lengths from the distribution of the observed projected lengths. A mixture of two two-parameter Weibull distributions has five parameters in total, which completely determines the mixed distribution. This is similar to the case that a normal distribution is determined by two parameters: its mean and its standard deviation. Now the second question arises as to how to convert the five parameters that determine the mixture distribution of projected lengths to the parameters of the mixture distribution of the original lengths. We will use the partial least squares (PLS) regression method to convert the parameters. Once the parameters of the original fiber length are obtained, its distribution is determined completely, and thus as a consequence various cotton quality parameters can be obtained directly from the distribution. Our calculations show that the PLS regression performs well. We believe that the method established in this study can help the cotton industry better understand the distribution of fiber length.

1.3 Definitions of parameters and instrument

1.3.1 Instrumentation in cotton classification

Figure 1.1: High Volume Instrument

Representatives of merchants and spinners throughout the world agreed that an international agreement on the use of instrument based quality evaluation systems is needed to standardize quality test results. The High Volume Instrument (HVI) was established to be the standard method for classification of fiber. See Figure1.1.

1.3.2 Definition of some quality parameters

Formal definitions of the following quality parameters are given in Section 3.1 of Chapter 3. A graphical representation of the parameters is given below using a schematic fibrogram of cotton. See Figure1.2.

1. Two expressions of mean or average length are available. One is mean length by weight of fibers; the other is mean length by number of fibers. The mean length by number of

fibers is formed at the intercept to the length axis by a tangent drawn from the origin of the Fibrogram at the amount axis. The mean length by weight of fibers is twice the area under the Fibrogram curve when the amount axis is normalized to unity (1.0 instead of 100%) (Spinlab 1981).

2. Upper half mean length: The average length of the longer half of the fibers.

3. Length uniformity index: The ratio between the mean length and the upper half mean length of the fibers, expressed as a percentage.

4. Short fiber content: The percentage of fibers in a sample, by weight, less than one half inch in length (Bargeron, 1991). Direct short fiber content measurements can be made with methods such as the Suter-Webb Array and AFIS. Another option for obtaining a measurement of short fiber is through the HVI system.

1.3.3 **Materials and method**

Eight sets of cotton samples of different lengths with the UHML 0.94 to 1.19 inches were selected for preparing fiber beards by using an HVI Fibrosampler. The fiber beards were collected from the clamp and tested on an AFIS. Four types of length distributions that are related to HVI measurements were measured by using AFIS:

1. Length distribution of the original sample, which was randomly selected by hand in small pinches from the sample population.

2. Length distribution of fibers sampled by the HVI Fibrosampler clamp.

3. Length distribution of fibers projecting from the clamp, which is the portion that is actually measured by the instrument using a beard method.

4. Length distribution of the hidden portion of fibers held in the clamp (invisible for an instrument to measure using a beard method).

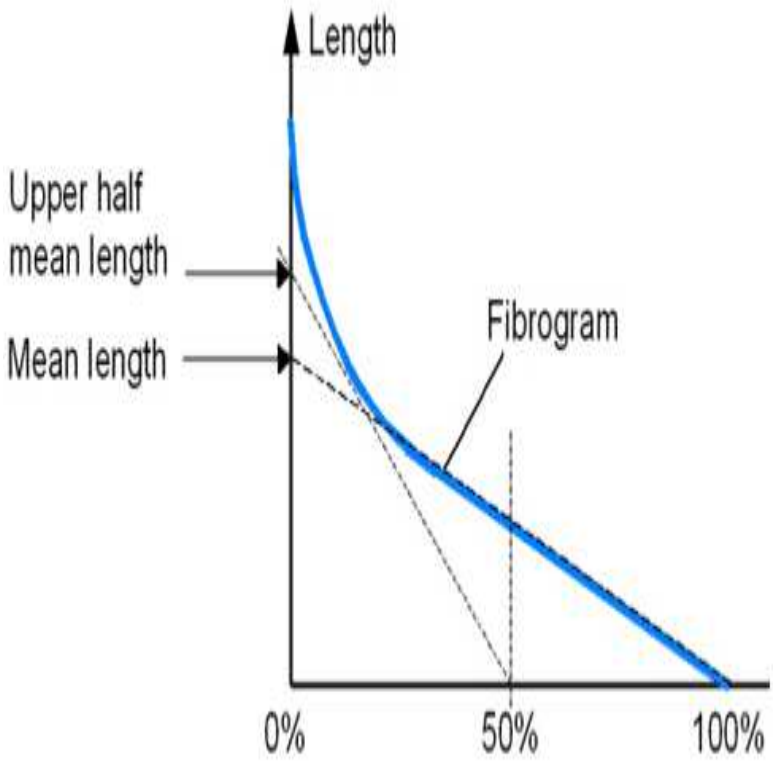

Figure 1.2: Schematic Fibrogram of Cotton

The projecting fibers were cut off along the baseline of the HVI clamp as shown in Figure1.3 and the projecting fibers were spray-dyed to show the hidden portion. In this study the first three types of length distributions are discussed.

Figure 1.3: Projecting and hidden portions of a beard from the HVI clamp

The frequency-length relationship of the above four types of lengths were used to construct the probability density functions (PDF) for fiber length distributions by number and by weight. Figure 1.4 shows the PDFs by number of the above length distributions of one cotton sample.

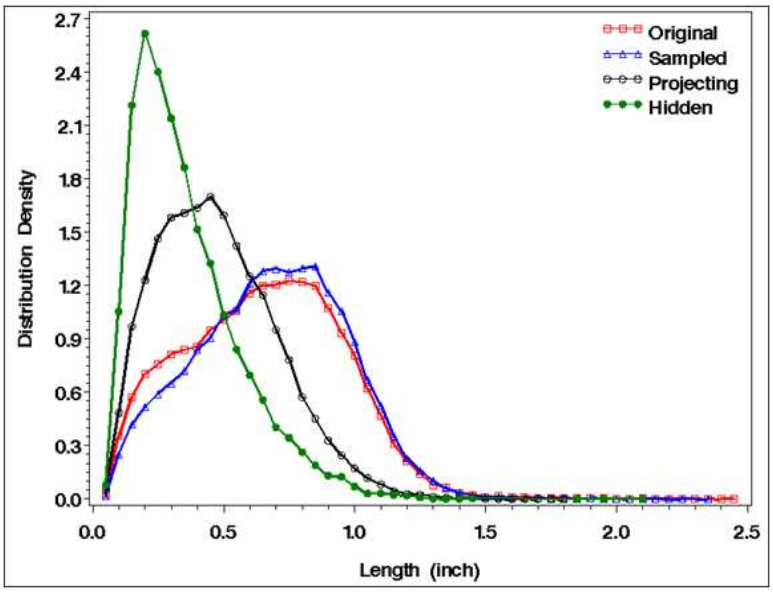

Figure 1.4: PDFs by number of the original sample, fibers picked by the HVI clamp, projecting portion of fibers, and the hidden portion of fibers.

As can be seen the measurement results (frequency-length relationship) of the four types of lengths were used to construct the probability density functions for fiber length distributions by number and by weight.

2

Chapter 2

Obtaining the Distribution of Actual Cotton Fiber Length By Number

2.1 Theoretical distribution of fiber length

2.1.1 Introduction

In this section non-linear regression models were constructed with different theoretical distributions. Gauss-Newton algorithm and least squares principle were used to solve the models and search for the PDFs that match the PDFs of the measured data. Statistical software SAS was utilized for the computational analysis. The results showed that a mixture of two two-parameter Weibull distributions is in good agreement with the available data. That is, each sample can be characterized by a mixture of Weibull distributions. Each mixed distribution has five parameters which may vary from sample to sample but similar samples have similar parameters.

1. **Definition of cumulative distribution function**

   The cumulative distribution function (CDF) completely describes the probability distribution of real-valued random variable X. For every real number x, the CDF of X is given by

$$x \mapsto F_X(x) = P(X \leq x)$$

where the right-hand side represents the probability that the random variable X takes on a value less than or equal to x. The CDF of X can be defined in terms of the probability density function f as follows:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

2. Empirical distribution function

Let $X_1, X_2, \ldots X_n$ be independent and identically distributed (iid) random variables with CDF $F(x)$. The empirical distribution function $F_n(x)$ based on sample $X_1, X_2, \ldots X_n$ is a step function define by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(X_i \leq x)}. \tag{2.1}$$

where I_A is the indicator of event A.

It has been shown that the empirical cumulative distribution function (ECDF) F_n is the non-parametric maximum likelihood estimate of the true CDF F and that the ECDF converges to F with probability one. That is, the probability of $\lim_{n \mapsto \infty} F_n(x) = F(x)$ equals one. In other words, when the sample size n goes to infinity, the ECDF converges to the true CDF F. Since the sample sizes of the data sets used in this project are considerably large ($> 30,000$), the ECDFs based upon these data sets can be considered as the underlying population CDFs of corresponding data sets. These ECDFs were the target to fit the distribution functions desired.

3. **Non-parametric least squares estimate**

The probability density function $g^*(x)$ is the least squares (LS) estimate of PDF $f^*(x)$ if $\int_0^\infty [g(x) - f^*(x)]^2 dx$ is minimized at $g^*(x)$ among all choices of $g(x)$, where x represent the fiber length.

2.1.2 Methodology

Cotton fiber length data was used to fit a distribution by initially dividing the interval $[0, 3]$ into 30 subintervals of equal lengths. The interval $[0, 3]$ was selected since it mainly covers the entire possible length of cotton fibers. Based on the selected interval, the frequencies $h(x)$ was counted (i.e. the number of fibers with length falling into a subinterval). That is, $h(x)$ equals the number of fibers with length falling into the subinterval covering x. in this equation; $h(x)$ represents the PDF of the cotton fiber length by number and is used as the underlying PDF of the data which at the same time is the $f^*(x)$ in the LS estimation process mentioned above. After a various attempts of different distributions including normal, lognormal, beta, Weibull, mixture of normal distributions, etc., it was found that a mixture of two Weibull distributions is considered the optimum solution.

1. **Definition of Weibull distribution function**

The PDF of a two-parameter Weibull distribution is given by

$$f(x, \lambda, \theta) = \lambda\theta x^{\lambda-1} e^{-\theta x^\lambda}. x > 0, \lambda > 0, \theta > 0$$

where λ and θ are parameters. The CDF is given by

$$F(x, \lambda, \theta) = \int_0^x f(x, \lambda, \theta) dt = 1 - e^{-\theta x^\lambda}, x > 0$$

2. **Definition of mixture of Weibull PDFs**

The PDF of a mixture of two Weibull PDFs is given by

$$f(x;\alpha,\lambda_1,\theta_1,\lambda_2,\theta_2) = \alpha f_1(x;\lambda_1,\theta_1) + (1-\alpha) f_1(x;\lambda_2,\theta_2) \tag{2.2}$$

where $0 < \alpha < 1$ and $f_i(x,\alpha,\lambda_i,\theta_i)$ is the PDF of a Weibull distribution, $i = 1, 2$. Therefore, the PDF of the mixture of two Weibull PDFs contains five parameters $\alpha, \lambda_1, \theta_1, \lambda_2, \theta_2$. Similarly, the CDF of the mixture is

$$F(x,\alpha,\lambda_1,\theta_1,\lambda_2,\theta_2) = \alpha F_1(x;\lambda_1,\theta_1) + (1-\alpha) F_2(x;\lambda_2,\theta_2)$$

where $F_i(x;\lambda_i,\theta_i)$ is the CDF of $f_i(x;\lambda_i,\theta_i)$, $i = 1, 2$. $f(x)$ was used for $f(x;\alpha,\lambda_1,\theta_1,\lambda_2,\theta_2)$ and $F(x)$ for $F(x;\alpha,\lambda_1,\theta_1,\lambda_2,\theta_2)$ to simplify notations. As mentioned earlier, due to the large sample size, this empirical PDF $h(x)$ can approximate the underlying population PDF of the data. Therefore, the LS estimate $g^*(x) = f(x)$, mixture of Weibull distributions, can be used as the population PDF by number. Once the functional form of the population PDF is obtained, various cotton quality parameters can be obtained.

2.2 Goodness-of-fit test

Kolmogorov-Smirnov goodness-of-fit test was performed to verify that the mixture of two Weibull distributions does fit the data. This test can be explained as follow: Let the hypothesis be "the data follows distribution $G(x)$", where $G(x)$ is a completely specified CDF. Let $F_n(x)$ denote the ECDF of a data. Define

$$D_n = \sqrt{n} \sup_{-\infty<x<\infty} | F_n(x) - G(x) |, x > 0$$

It is shown that when the hypothesis is true and the sample size n is large (Mood, et al

1974), D_n is approximately distributed as $D(x)$, where

$$D(x) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2x^2}, x > 0$$

It is clear that if the hypothesis is false, then D_n tends to be large; hence for a given significance level α, one would reject the hypothesis if D_n is greater than the critical value d_α, where d_α is determined so that $D(d_\alpha) = 1 - \alpha$. For example, when $\alpha = 0.10$, $d_\alpha \simeq 1.22$, and when $\alpha = 0.05$, $d_\alpha \simeq 1.44$. Then, the hypothesis is rejected if D_n is greater than the critical value $d_\alpha \simeq 1.22$ at $\alpha = 0.10$, and it is rejected if D_n is greater than the critical value $d_\alpha \simeq 1.44$ at $\alpha = 0.05$.

The Kolmogorov-Smirnov goodness-of-fit test is performed in the following steps:

1. Since in practice it is common that the sample size is usually around 2000 to 3000, we randomly re-sample n=2500 observations from a data set.

2. Fit a mixture of two Weibull distributions to the re-sampled data set. This fitted mixture CDF is $G(x)$, and the ECDF of the re-sampled data is $F_n(x)$ in Equation 2.1.

3. Use $\alpha = 0.10$. Compare D_n with $d_\alpha \simeq 1.22$. If D_n is less than 1.22, the Kolmogorov-Smirnov statistic is not significant, then the hypothesis that these 2500 re-sampled data points follow a mixture of two Weibull distributions is accepted.

4. Repeat Steps 1 to 3 500 times and record the number of times that the hypothesis is accepted. In all the eight sets of fibers, the number of acceptance is higher than 95% of the 500 tests for each set.

The test was performed for original, sampled, and projecting fibers of all eight cottons, and the hypothesis was accepted in all cases with a p-value greater than 0.1. Therefore, the fiber length distribution by number can be described using a mixture distribution of two Weibull distributions.

ID	Type	Rep	Pct $d_\alpha \simeq 1.22$	Pct $d_\alpha \simeq 1.44$	Std at $d_\alpha \simeq 1.22$	Std at $d_\alpha \simeq 1.44$
30	Original	500	1	1	0	0
	Sampled	500	1	1	0	0
	Projecting	500	0.97	0.99	0.00764	0.00445
31	Original	500	1	1	0	0
	Sampled	500	1	1	0	0
	Projecting	500	0.996	0.998	0.00283	0.002
33	Original	500	1	1	0	0
	Sampled	500	1	1	0	0
	Projecting	500	0.998	1	0.002	0
34	Original	500	1	1	0	0
	Sampled	500	1	1	0	0
	Projecting	500	0.994	0.994	0.00346	0.00346
35	Original	500	1	1	0	0
	Sampled	500	1	1	0	0
	Projecting	500	0.984	0.99	0.00562	0.00445
36	Original	500	1	1	0	0
	Sampled	500	1	1	0	0
	Projecting	500	0.998	1	0.002	0
37	Original	500	1	1	0	0
	Sampled	500	1	1	0	0
	Projecting	500	0.992	1	0.00399	0
38	Original	500	1	1	0	0
	Sampled	500	1	1	0	0
	Projecting	500	1	1	0	0

Table 2.1: Kolmogorov-Simirnov Goodness of Fit Test

The first column of the above table represents the mean length by number of the eight samples used in the study. In particular, Sample ID 30 denotes the cotton sample with mean length of 30/32 inches, and ID 31 denotes the cotton sample with mean length of 31/32 inches, etc. Three types of fiber lengths, original, sampled and projecting, are presented in the second column. The third column is the number of repeatations in our re-sampling with replacement. The fourth and the fifth columns give the percentage of cases where the fiber length distribution by number can be described by a mixture of two Weibull distributions at two different levels $d_\alpha \simeq 1.22$ at $\alpha = 0.10$, and $d_\alpha \simeq 1.44$ at $\alpha = 0.05$. For example, the first row shows 100% of the 500 samples are accepted, and the third row shows 97% of the 500 samples are accepted and so on. Finally, the last two columns show how much variation or error we have in our test.

2.3 Estimation of mixture distribution parameters by number and by weight

The above discussion is regarding the fiber length distribution by number. The shorter fibers may have a large number portion, but not a large weight portion. Therefore, in practice, the cotton fiber length distribution by weight is more commonly used. The fiber length has a very weak correlation with the fiber linear density, therefore, the fiber length and fiber linear density are independent. As previously defined, $h(x)$ denotes the frequency function of a data set and let $\overline{x}$ denote the sample mean of the data set. Assuming the independence between fiber length and fiber linear density, the frequency function by weight is given by (Zeidman et al, 1991)

$$h_w(x) = \frac{xh(x)}{\overline{x}} \tag{2.3}$$

A mixture of two Weibull distributions is used to fit $h_w(x)$, and Kolmogorov-Smirnov goodness-of-fit test is also performed for the length by weight and had the same conclusions as for the length by number. Table 2.1 and Table 2.2 present the parameters of the estimated mixture distributions by number and by weight for the eight different cottons, respectively. For simplicity, only the graphs of the PDF's by number and by weight for Sample ID 34 and Sample ID 38 are shown in Figure 2.1 to Figure 2.12 . It can be seen that both curves between the sample and the estimated are in good agreement. (λ_1, θ_1) are the parameters of the Weibull distribution on the left (labeled pdf1 in figures 2.1 to 2.12) of the two Weibull distributions in the mixture, which represents mainly shorter fibers, and the one on the right determined by (λ_2, θ_2) represents longer fibers (labeled pdf2 in figures 2.1 to 2.12).

ID	Type	α	λ_1	θ_1	λ_2	θ_2
30	Original	0.229	2.114	1.336	3.481	0.073
	Sampled	0.502	1.980	0.390	4.283	0.032
	Projecting	0.946	2.325	0.465	3.383	5.640
31	Original	0.840	3.667	0.060	2.197	1.879
	Sampled	0.166	2.143	1.204	3.752	0.055
	Projecting	0.039	3.129	4.471	2.367	0.407
33	Original	0.301	1.881	0.910	4.042	0.026
	Sampled	0.714	4.006	0.027	1.958	0.614
	Projecting	0.031	3.055	4.033	2.335	0.342
34	Original	0.508	5.178	.006	1.841	0.338
	Sampled	0.525	5.032	.007	2.076	0.277
	Projecting	0.060	2.960	3.082	2.449	0.292
35	Original	0.515	4.931	.007	1.921	0.332
	Sampled	0.468	4.924	.008	2.014	0.303
	Projecting	0.934	2.358	0.296	2.723	2.149
36	Original	0.554	4.971	.005	1.674	0.416
	Sampled	0.591	4.612	.009	1.995	0.340
	Projecting	0.090	2.370	1.634	2.369	0.280
37	Original	0.623	4.632	.007	1.738	0.491
	Sampled	0.558	4.884	.005	2.034	0.275
	Projecting	0.057	2.527	1.820	2.302	0.276
38	Original	0.645	5.151	.003	1.779	0.502
	Sampled	0.546	5.396	.002	1.942	0.278
	Projecting	0.099	2.245	1.404	2.488	0.206

Table 2.2: Estimation of Mixture Distribution Parameters by Number

The above table presents the five parameters of the mixture of Weibull distributions by number as given in Equation (2.2).

ID	Type	α	λ_1	θ_1	λ_2	θ_2
30	Original	0.531	4.828	0.017	2.356	0.181
	Sampled	0.481	5.017	0.014	2.731	0.124
	Projecting	0.820	2.751	0.224	2.640	0.387
31	Original	0.706	4.782	0.018	2.116	0.196
	Sampled	0.728	4.606	0.021	2.248	0.180
	Projecting	0.721	2.828	0.218	2.899	0.206
33	Original	0.618	5.188	0.006	2.252	0.175
	Sampled	0.666	4.817	0.009	2.532	0.143
	Projecting	0.842	2.863	0.147	3.145	0.261
34	Original	0.607	5.653	0.003	2.373	0.123
	Sampled	0.181	2.748	0.224	5.127	0.006
	Projecting	0.998	2.891	0.152	5.482	5.723
35	Original	0.524	5.846	0.002	2.632	0.096
	Sampled	0.550	2.770	0.087	5.902	0.003
	Projecting	0.998	2.758	0.151	419.980	9.9E-9
36	Original	0.430	2.456	0.104	6.087	0.001
	Sampled	0.512	5.846	0.002	2.835	0.075
	Projecting	0.996	2.776	0.148	8.161	0.480
37	Original	0.386	2.358	0.114	5.682	0.002
	Sampled	0.539	5.960	0.001	2.836	0.070
	Projecting	0.999	2.750	0.137	107.020	145.996
38	Original	0.342	2.295	0.116	6.182	0.001
	Sampled	0.575	6.410	0.001	2.723	0.073
	Projecting	0.998	2.872	0.110	11.964	10E-7

Table 2.3: Estimation of Mixture Distribution Parameters by Weight

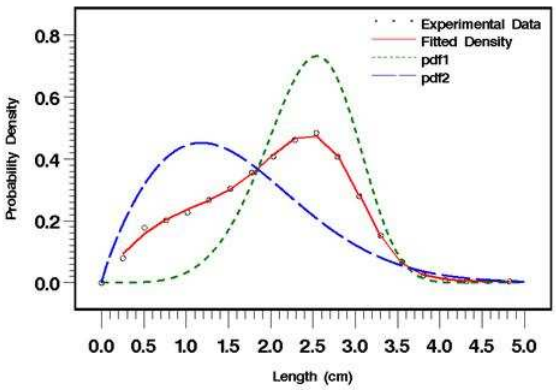

Figure 2.1: Probability density functions by number of ID 34 original fibers.

looking at the graph from left to right,the thick-dashed line and thin-dashed line represent, respectively, the two pdf's of the two Weibull distributions. The continuous line represents the mixture of the two distributions. The experimental data is represented by the dotted curve. Hence, the graphical relationships between the PDF of the estimated mixture distributions and the PDF of experimental data are in good fit for the ID 34 original fiber length by number, which was randomly selected by hand in small pinches from the sample population.

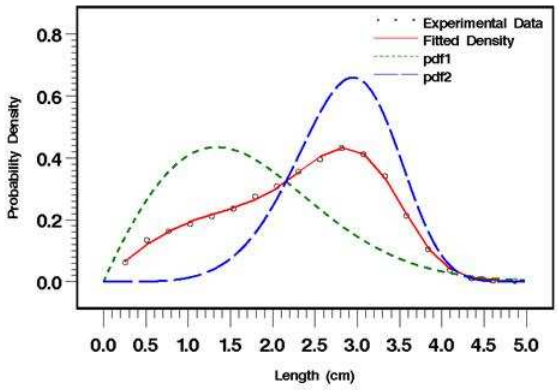

Figure 2.2: Probability density functions by number of ID 34 HVI sampled fibers.

This figure shows the comparison between the PDF of the estimated mixture distributions and the PDF of the experimental data by number. This experiment was performed for observation 34 that was sampled by the HVI Fibrosampler clamp.

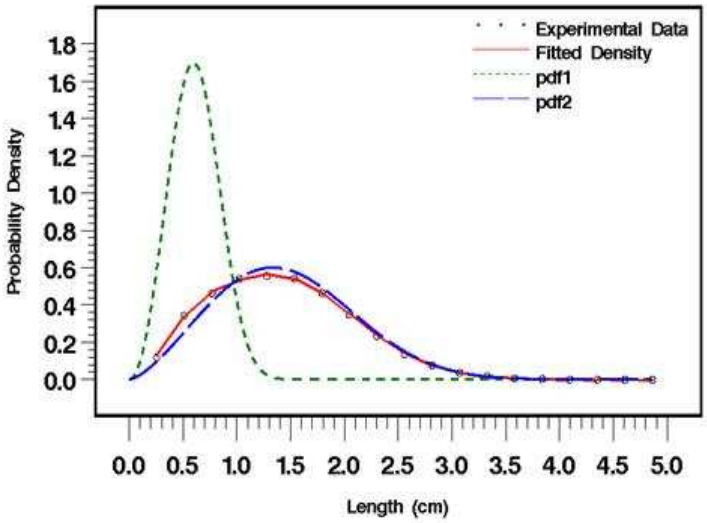

Figure 2.3: Probability density functions by number of ID 34 projecting fibers.

This figure shows the comparison between the PDF of the estimated mixture distributions and the PDF of the experimental data by number. This experiment was performed for ID 34 projecting fibers which is the portion that is actually measured by the HVI instrument using a beard method.

Similarly, the rest of the graphs show the observations ID 34 and ID 38 original fiber length, sampled fiber length, and projecting fiber length by number and by weight.

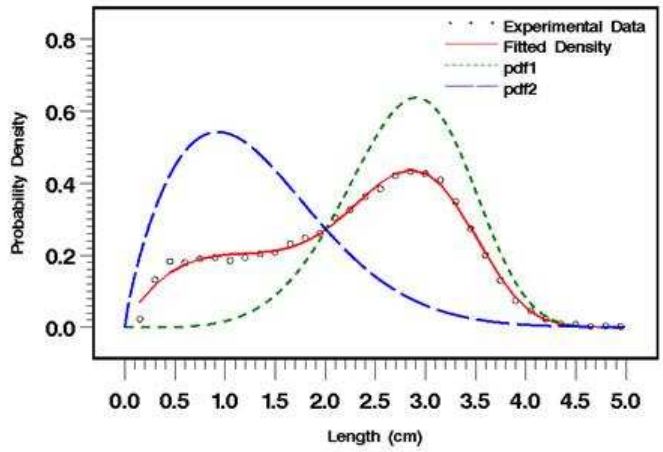

Figure 2.4: Probability density functions by number of ID 38 original fibers.

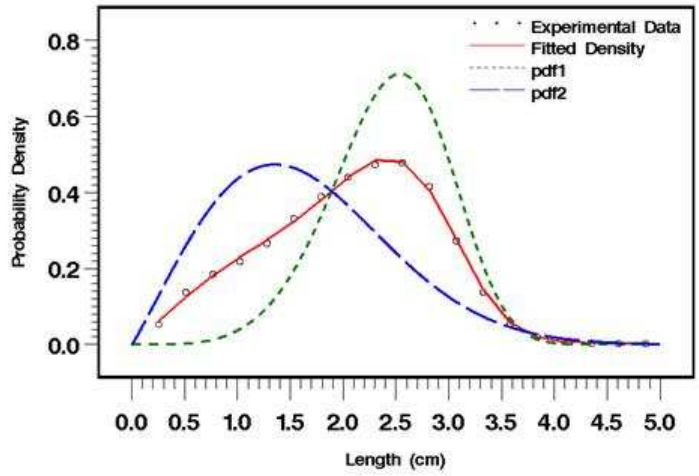

Figure 2.5: Probability density functions by number of ID 38 HVI sampled fibers.

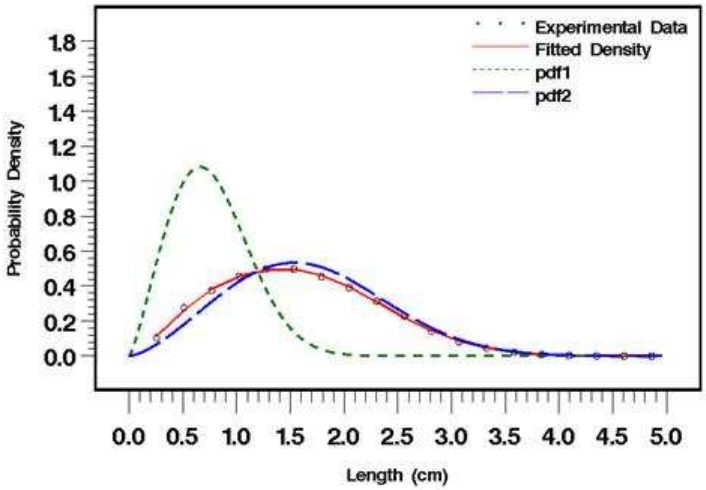

Figure 2.6: Probability density functions by number of ID 38 projecting fibers.

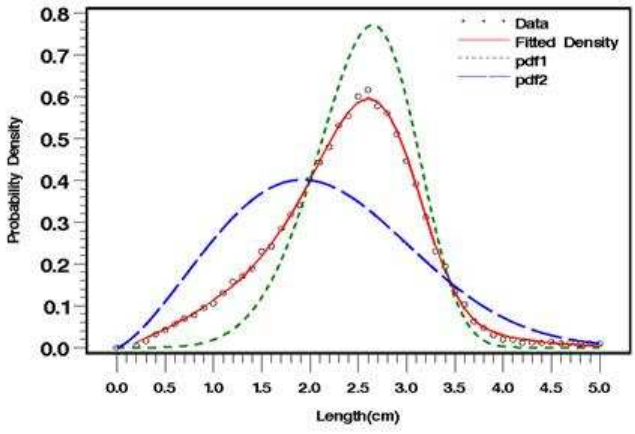

Figure 2.7: Probability density functions by weight of ID 34 original fibers.

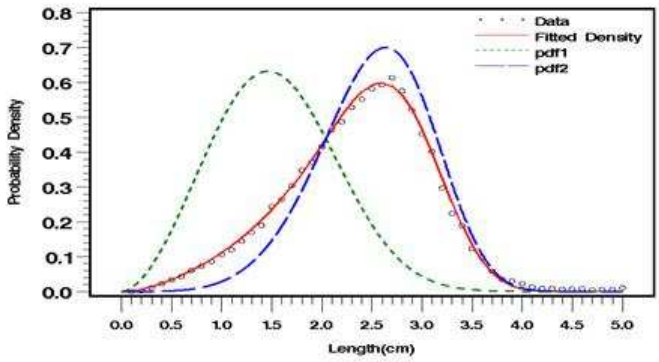

Figure 2.8: Probability density functions by weight of ID 34 HVI sampled fibers.

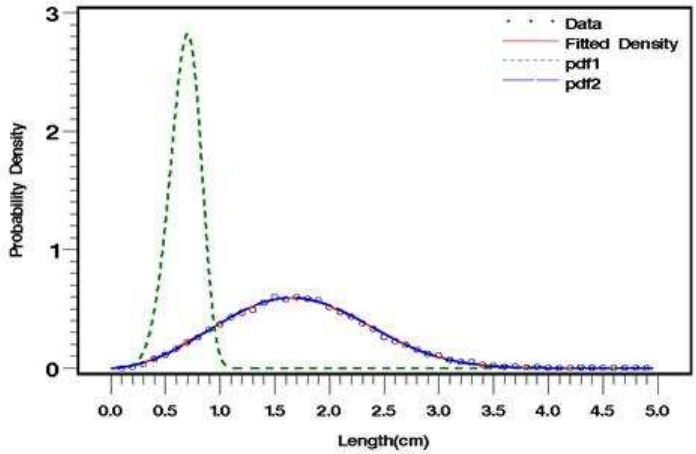

Figure 2.9: Probability density functions by weight of ID 34 projecting fibers.

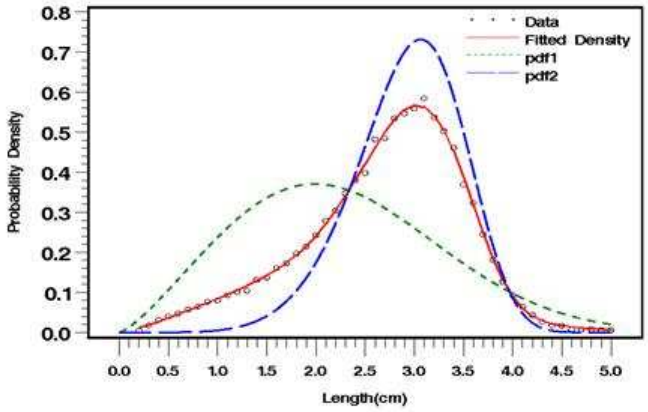

Figure 2.10: Probability density functions by weight of ID 38 original fibers.

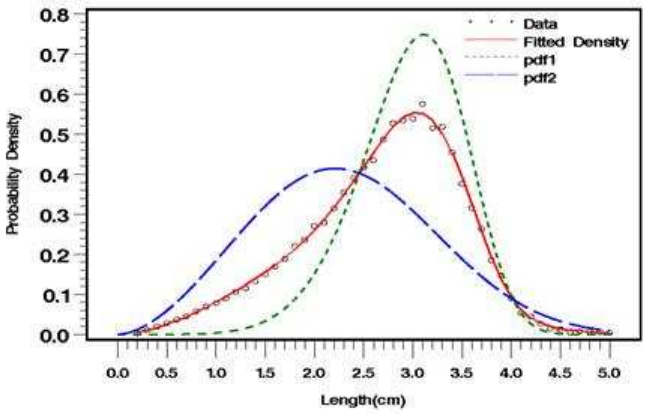

Figure 2.11: Probability density functions by weight of ID 38 HVI sampled fibers.

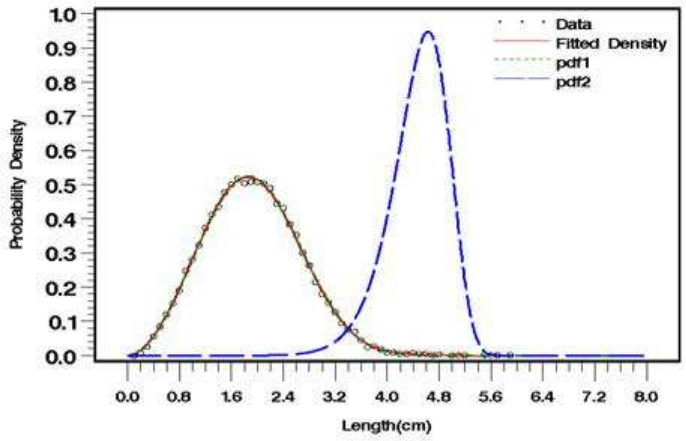

Figure 2.12: Probability density functions by weight of ID 38 projecting fibers.

2.4 Conclusion

We studied the theoretical distribution functions that can describe the underlying distributions of three types of fiber lengths that are related to HVI measurements: the lengths of the original fiber population, of the fibers picked by HVI fibrosampler, and of the beard's projecting portion that is actually scanned by HVI.

We conducted non-linear regressions based on length data measured by using AFIS from eight cottons. A mixture of two Weibull distributions fits the data very well. Kolmogorov-Smirnov goodness-of-fit test confirms that a mixed Weibull distribution can be used as the underlying distribution of fiber length. The goal is to find a model to predict approximately the original real length (Fiber collected from the HVI clamps that is not cut), since we cannot measure the hidden portion of the fiber. Now, while the distribution of fiber length is a mixture of two Weibull distributions, which is determined by five parameters, conversion from the distribution function of projected lengths to that of the original lengths has become a conversion from the parameters that determine the mixture distribution of projected lengths to those of the mixture distribution of the original lengths. A natural choice of the conversion method is ordinary least squares regression (OLS). However, OLS is not a good method for our data because of the small sample size, which is eight (eight cotton samples), and a high collinearity (multicollinearity) between variables (columns in Table 2.2). Therefore, to overcome these problems, we are using partial least squares (PLS) regression for the distribution parameter conversion. The highlight of some characteristics and steps of PLS are presented in Chapter 4.

3

Chapter 3

Quality Parameters

3.1 Estimation of some quality parameters

3.1.1 Fiber length parameters from the mixture of Weibull distributions

In this section, comparisons between the parameters obtained from the original data and that obtained from the estimated mixture distribution are presented. The purpose of this chapter is to investigate the "closeness" of the estimation. These parameters include upper half mean length (UHML), coefficient of variance (CV), short fiber content (SFC), uniformity index (UI), and recently introduced lower half mean length (LHML) (Cui et al., 2004).

The mean and the variance of a Weibull distribution with parameters λ and θ are, respectively, given by

$$\mu = \Gamma(1 + 1/\lambda)\theta^{1/\lambda} \tag{3.1}$$

and

$$\sigma^2 = [\Gamma(1 + 2/\lambda) - \Gamma^2(1 + 1/\lambda)]\theta^{-2/\lambda} \tag{3.2}$$

where $\Gamma(x) = \int_0^\infty e^{-x}dx$ denotes the gamma function. Hence, the mean and the variance of a mixed Weibull distribution are, respectively,

$$\mu_{mix} = \alpha\mu_1 + (1-\alpha)\mu_2. \tag{3.3}$$

and

$$\sigma_{mix} = E_{mix}(x^2) - \mu_{mix}^2 = \alpha(\sigma_1^2 + \mu_1^2) + (1-\alpha)(\sigma_2^2 + \mu_2^2) - [\alpha\mu_1 + (1-\alpha)\mu_2]^2 \tag{3.4}$$

where μ_i and σ_i denote, respectively, the mean and the variance of a Weibull distribution with parameters λ_i and θ_i, $i = 1, 2$.

The mean fiber length can be obtained by $\mu = \int_0^L [1-F(u)]du$, where L is the maximum fiber length in the distribution. The Fibrogram is a curve defined by $B(x) = \frac{1}{\mu}\int_x^L [1-F(u)]du$ (Zeidman et al, 1991). The UHML is the mean length of those fibers that are longer than the median value of the weight distribution of fibers. Let M denote the weight median value, such that $\int_0^M uf(u)du = \frac{\mu}{2}$. It is known that the UHML is the x-intercept of the tangent line to the Fibrogram $B(x)$ passing through y-intercept 0.5, and can be obtained by:

$$UHML = \frac{\int_M^L uf(u)du}{1-F(M)} \tag{3.5}$$

The LHML is given by

$$LHML = \frac{\int_0^M uf(u)du}{F(M)} \tag{3.6}$$

The SFC by number is the number proportion of fibers shorter than 0.5 inches in the cotton. When using the fitted distribution, the SFC by number is given by:

$$\alpha F_1(0.5, \lambda_1, \theta_1) + (1-\alpha)F_2(0.5, \lambda_2, \theta_2). \tag{3.7}$$

Similarly, the SFC by weight can be defined. As pointed out in (Spinlab 1981), (Hertel 1940), the UHML is chosen purely for convenience, since it can be obtained by drawing a tangent line. The interpretation of the UHML is actually a proportion of fibers by number longer than the median value of fibers by weight.

3.1.2 **Application and comparison**

To make numerical comparisons between the length distribution from the data and that from the estimated mixed Weibull disrbution, we compare some most important quality parameters used in the cotton industry. Plugging the fitted mixed Weibull distribution into formulas (3.1)-(3.7) will give us the estimated mean length, UHML, LHML, SFC and CV. We present the results in Tables 3.1 and 3.2 below. Specifically, for each cotton ID in these two tables there are three type of fibers, Original, Sampled or Projecting, and each fiber has two rows of values. For example, the top row of Original cotton ID 30 is from the experimental data and the second row is from the fitted mixture Weibull distribution. It can be seen from Table 3.1 and Table 3.2, that the parameter obtained from the experimental data and that obtained from the mixed Weibull distribution match very well. Considering the natural non-uniformity of cotton length, the third decimal of the data usually does not have statistical significance. If we round up the data from Tables 3.1 and 3.2 to the second decimal, the results obtain from tests and from estimation are identical thus indicating an excellent fit of the mixture of the Weibull distributions.

ID 30	Mean	UHML	LHML	CV	SFC	UI
Original	1.666	2.497	1.250	0.453	0.318	80.42
	1.647	2.466	1.237	0.448	0.323	80.25
Sampled	1.740	2.511	1.331	0.415	0.271	81.23
	1.728	2.502	1.322	0.425	0.276	81.32
Projecting	1.210	1.892	0.889	0.480	0.577	78.68
	1.194	1.863	0.879	0.478	0.579	78.74
ID 31						
Original	1.754	2.531	1.342	0.423	0.261	81.67
	1.738	2.504	1.331	0.418	0.266	81.56
Sampled	1.782	2.525	1.377	0.401	0.245	81.91
	1.767	2.499	1.366	0.396	0.249	81.79
Projecting	1.271	1.951	0.942	0.464	0.530	79.17
	1.267	1.949	0.938	0.466	0.531	79.07
ID 33						
Original	1.873	2.807	1.405	0.458	0.267	80.74
	1.845	2.764	1.384	0.456	0.275	80.64
Sampled	1.937	2.786	1.484	0.414	0.221	81.42
	1.920	2.764	1.471	0.413	0.227	81.34
Projecting	1.385	2.137	1.025	0.468	0.465	79.04
	1.377	2.125	1.018	0.469	0.467	79.04
ID 34						
Original	2.041	2.928	1.567	0.419	0.208	81.94
	2.043	2.943	1.565	0.422	0.211	81.76
Sampled	2.071	2.907	1.608	0.392	0.182	82.21
	2.074	2.915	1.609	0.392	0.186	82.06
Projecting	1.424	2.184	1.057	0.464	0.440	79.28
	1.415	2.170	1.049	0.463	0.443	79.14

Table 3.1: Estimation of Some Quality Parameters by Number of 30 till 34

ID 35	Mean	UHML	LHML	CV	SFC	UI
Original	2.058	2.967	1.575	0.419	0.206	81.54
	2.056	2.971	1.572	0.420	0.210	81.42
Sampled	1.999	2.881	1.530	0.415	0.214	81.35
	1.997	2.889	1.526	0.418	0.218	81.22
Projecting	1.439	2.237	1.061	0.476	0.440	78.89
	1.431	2.229	1.054	0.476	0.444	78.79
ID 36						
Original	2.145	3.132	1.631	0.436	0.208	81.53
	2.149	3.167	1.626	0.447	0.215	81.40
Sampled	2.151	3.074	1.654	0.408	0.185	81.63
	2.122	3.028	1.634	0.406	0.188	81.62
Projecting	1.456	2.272	1.071	0.479	0.435	78.82
	1.445	2.262	1.062	0.481	0.439	78.67
ID 37						
Original	2.176	3.194	1.650	0.441	0.205	81.42
	2.168	3.189	1.642	0.442	0.211	81.30
Sampled	2.255	3.203	1.740	0.403	0.165	81.87
	2.253	3.201	1.738	0.402	0.168	81.75
Projecting	1.508	2.358	1.108	0.482	0.411	78.77
	1.501	2.352	1.103	0.482	0.414	78.69
ID 38						
Original	2.284	3.301	1.746	0.432	0.192	82.07
	2.298	3.335	1.753	0.435	0.197	81.96
Sampled	2.324	3.293	1.796	0.404	0.164	82.11
	2.371	3.388	1.824	0.412	0.166	81.89
Projecting	1.588	2.447	1.176	0.466	0.368	79.02
	1.584	2.447	1.171	0.469	0.371	79.01

Table 3.2: Estimation of Some Quality Parameters by Number of 35 till 38

3.2 Conclusion

The length parameters, such as mean length and upper half mean length calculated from the mixture of Weibull distributions match extremely well with those calculated from the actual data. Since the distribution of fiber length can be described as a mixture of two Weibull distributions, which in turn is determined by five parameters, the relationship between the length distribution of projecting fibers and that of the original fibers can be investigated by exploring the relationship between the five parameters of the mixture Weibull distributions and the five parameters of projecting fiber length distributions.

4

Chapter 4

Theory of Partial Least Squares Regression

4.1 Partial Least Squares

The partial Least Squares (PLS) regression method is of vitel importance in many fields. The original work in PLS was done in the late 60's by Herman Wold in the field of econometrics. The use of the PLS method was pioneered by Svante Wold and Harald Martens in the late 70's. The goal of PLS is to predict the dependent variables from the independent variables and to describe the common structure underlying the two variables (Abdi, 2003). It is illustrated that PLS is a better tool than the classical OLS regression because the former is more robust. This chapter briefly describes the highlight of PLS. PLS regression is a multivariate regression that has some similarity with principal component regression (PCR) (Svante Wold 1986). The PCR first summarizes information from independent variables into principal components and then constructs regression equation between the principal variables which are new independent variables and the original dependent variables. Since the number of principal components is usually less than the original number of independent variables, the PCR can overcome some difficulties that the original ordinary regression may have. The difference between PLS and PCR is that PCR uses summarized information, from independent variables alone but PLS utilizes summarized information, called latent variable or factor, from both independent and dependent variables. Similar to principal components

the factors extracted are orthogonal. Once the factors are obtained, an ordinary least squares regression of original dependent variables against factors is used for prediction.

Let $y_i = (y_{1i}, ..., y_{ni})^T$ denote a vertical vector of observations of dependent variable Y_i, $1 \leq i \leq l$, and $x_j = (x_{1j}, ..., x_{nj})^T$ denote a vertical vector of observations of independent variable X_j, $1 \leq j \leq m$, and $((x_{i1}, \ldots, x_{im})^T, (y_{i1}, \ldots, y_{il})^T)$, $i = 1, 2, \ldots, n$ be paired observations, where A^T denotes the transpose of matrix A. Furthermore, let $D = (x_1, \ldots, x_m)$ denote the observed independent variable matrix, and $Q = (y_1, \ldots, y_l)$ denote the observed dependent variable matrix.

Unlike OLS regression or PCR, PLS regression is designed to overcome problems such as small sample size and/or multicollinearity between variables. PLS is more effective than OLS when there is a high collinearity or multicollinearity between the independent variables, the response variables, or both of them, and the sample size is small. PLS can even be used when the sample size is less than the number of independent variables, which of course does not work for OLS regression. generally speaking, PLS regression searches for a set of factors that performs a simultaneous decomposition of D and Q with the constraint that these factors explain the maximum covariance between D and Q. In the univariate case, i.e., there is only one dependent variable, the PLS regression is sometimes called PLS1, while in the multivariate case, i.e., when there are two or more dependent variables, it is sometimes called PLS2 (Garthwaite 1994).

For our data, the independent variable vector is $(a', \lambda_1', \theta_1', \lambda_2', \theta_2')^T$ representing the five parameters of the mixed Weibull distribution from the projecting fiber length. The dependent variable vector is $(a, \lambda_1, \theta_1, \lambda_2, \theta_2)^T$ representing the parameters of the mixed Weibull distribution from the actual fiber length. Thus $l = m = 5$. Since there are eight cottons in our study which generates five parameters from the projecting length and five parameters from the original length, eight pairs of observations were obtained. That is, sample size $n = 8$, which is considered small. Meanwhile, notice that in addition to small

sample size, components of independent variables in our data are obviously highly correlated as well as the dependent variables. Therefore, the use of OLS regression in such case is not appropriate, thus PLS regression is considered. The objective of PLS regression, like OLS regression, is to establish an equation between dependent variables and independent variables such that once the parameters of the distribution of projecting length are obtained, one can use the equation to obtain the parameters of the distribution of the corresponding original length.

4.2 Theory of PLS

4.2.1 Theory of PLS1

Garthwaite (1994) gave an excellent description on the fundamental ideas of PLS, formulas of obtaining factors, and ways to determine the number of factors desired. In this section we briefly illustrate the approach of PLS1.

Without loss of generality we assume that D and Q are centered and scaled. The objective is to obtain the following regression equation

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 T_1 + \hat{\beta}_2 T_2 + ... + \hat{\beta}_p T_p. \tag{4.1}$$

where p is the number of factors, and T_k is a factor, a linear combination of X_i, $1 \leq i \leq m$.

To find the first factor T_1, we use a least squares regression of Y against X_1 then against X_2 until X_m to have m simple regressions. Since we assumed that the independent variables are centered, the obtained least squares regression equations have no intercept. We denote the LS regression equations by

$$Y_{1j} = b_{1j} X_j, j = 1, \ldots, m,$$

where

$$b_{1j} = (x_j^T x_j)^{-1} x_j^T y.$$

Considering a weighted average of the above m regression equations and let

$$T_1 = \sum_{j=1}^{m} \omega_{1j} b_{1j} X_j,$$

where $\omega_{1j} \geq 0$ and $\sum_{j=1}^{m} \omega_{1j} = 1$. One choice of weight ω_j is $\omega_j = 1/m$ for all j. Another possible choice is $\omega \sim var(X_j)$ (Garthwaite 1994). Since each of the m regression equations contains certain amount of information about the dependent variable, so does T_1. Obviously, T_1 contains information not only about Y but about each independent variable X_j as well. This T_1, a linear combination of X_j's, is the first factor. We now consider the second factor and hope that the second factor can explain variation of Y and X that has not been explained by T_1.

Treat T_1 as an independent variable and let Y regress on T_1. Denote $r_1 = (r_{11}, \ldots, r_{1n})^T$ the corresponding residuals of the regression. Thus information of Y that is not explained by T_1 must be contained in r. Moreover, regress X_{1j} on T_1 and let x_{2j} denote the corresponding residuals of the regression for each $j = 1, 2, \ldots, m$. Similarly, information of X_j that is not explained by T_1 is contained in x_{2j}. Notice that since it is assumed that Y and X are centered and scaled, so are r_1, T_1, and x_{2j}. In order to account for the rest of variation that is not been explained by T_1, let R_1 denote the underlying random "error" with values r_1 and X_{2j} the underlying random "error" with values x_{2j}. Now regress residual R_1 against residual X_{2j} for each j and then like T_1, denote T_2 a linear combination of this second group of the m simple regressions. This T_2 is still a linear combination of X_j's and is the second factor desired. The weights in T_2 are not necessarily the same as those in T_1. Recall that in PCA the first principal component contains more information about X than all other principal components. Between T_1 and T_2, it is true that in terms of percentage T_1 contains more

information about Y and X than does T_2. It can be loosely stated that the information that T_1 contains and the information that T_2 contains are not overlapping. This is why and how T_1 and T_2 are örthogonal". Replacing T_1 with T_2, Y with R_1, and X_j with X_{2j}, one can repeat the steps of constructing T_2 to obtain T_3, T_4, ..., etc. In general, the amount of information of Y and X contained in T_k is more than that contained in T_{k+1}. There are different criteria to determine the number of factors needed (Garthwaite 1994). As mentioned earlier, once the factors are determined, treat them as independent variables and regress Y against them using OLS regression to have Equation (4.1).

4.2.2 **Theory of PLS2**

The objective of PLS2 is to determine a relation between the predictors and the responses in the multivariate level (Garthwaite 1994). In this section we are developing the multivariate PLS approach when response Y is a vector containing two or more variables. Let us keep the same notations as used in PLS1 such as $D = (x_1, \ldots, x_m)$, the observed independent variable matrix and $Q = (y_1, \ldots, y_l)$, the observed dependent variable matrix.

Assuming that D and Q are centered and scaled. Our goal is to obtain the following regression equation

$$\hat{Y_k} = \hat{\beta_{k0}} + \hat{\beta_{k1}} T_1 + \hat{\beta_{k2}} T_2 + ... + \hat{\beta_{kp}} T_p. \tag{4.2}$$

where $k = 1, \ldots, l$, p is the number of factors, and T_k, called factor, is a linear combination of X_j's.

The PLS2 consists of two steps:

1. The first step is to find the first factor based on the covariance between X and Y. Let C_1 be the eigenvector corresponding to the largest eigenvalue of the sample covariance matrix between X and Y, $Q'DD'Q$, and define $U_1 = QC_1$. Then the first factor T_1 is

obtained by following the procedures of PLS1 with Y (the Y in PLS1) replaced by U_1.

2. Then let each Y_i regress on T_1 and denote its residual by R_i, $1 \leq i \leq l$. Let $RY_2 = (R_1, \ldots, R_l)$ and Q_2 the matrix of the corresponding observed residual values (Note: Q_2 is similar to r in PLS1.). Furthermore, still like in PLS1, let X_j regress against T_1 for all j, X_{2j} denote its residual with $RX_2 = (X_{21}, \ldots, X_{2m})$ and $x_2 = (x_{21}, \ldots, x_{2m})$ be the observed residual values. Let C_2 be the eigenvector corresponding to the largest eigenvalue of the sample covariance matrix between residuals RY_2 and RX_2, which is $Q_2' x_2 x_2' Q_2$. By defining $U_2 = Q_2 C_2$, the second factor T_2 is then obtained by the procedures of obtaining T_1 with U_1 and X replaced by U_2 and RX_2 respectively. Similarly, one can obtain factors T_3, T_4, ..., by the same manner. Finally, Equation (4.2) is obtained by the OLS regression. We now provide an algorithm for obtaining Equation (4.2).

4.3 **PLS algorithm**

Let $\| v \|$ denote the norm of vector v. The above PLS2 procedures can be summarized by the following algorithm (Jrgensen and Goegebeur 2002). Set initial values $X_{<1>} = X$ and $Y_{<1>} = Y$.

1. Let $u_{<1>}$ be an arbitrary column of Y.
2. Let $w_{<1>} = \frac{X_{<1>}^T u_{<1>}}{\| X_{<1>}^T u_{<1>} \|}$.
3. Let $t_{<1>} = X_{<1>} u_{<1>}$ and $q_{<1>} = \frac{Y_{<1>}^T t_{<1>}}{\| Y_{<1>}^T t_{<1>} \|}$.
4. Redefine $u_{<1>} = Y_{<1>} q_{<1>}$.
5. If $u_{<1>}$ in Step 4 is the same as that in Step 1, then continue with Step 6; otherwise go back to Step 2 with the new $u_{<1>}$.

6. Let $c_{<1>} = \frac{t_{<1>}u_{<1>}}{t_{<1>}t_{<1>}}$.

7. Let $p_{<1>} = \frac{X_{<1>}^T t_{<1>}}{t_{<1>}^T t_{<1>}}$.

8. Finally, let $X_{<2>} = X_{<1>} - t_{<1>}p_{<1>}^T$ and $Y_{<2>} = Y_{<1>} - c_{<1>}t_{<1>}q_{<1>}^T$.

Then, repeat the above procedure by replacing $X_{<1>}$ and $Y_{<1>}$ with $X_{<2>}$ and $Y_{<2>}$ respectively, in steps 1 through 8 to find $u_{<2>}, w_{<2>}, t_{<2>}, c_{<2>}, q_{<2>}$, and $p_{<2>}$, etc. The number of factors desired can be determined by cross-validation among other criteria. Once the number of factors is determined, PLS regression factors and the final regression equation matrix B (explained in chapter 5) consist of the above established vectors $(u_{<i>}, w_{<i>}, t_{<i>}, c_{<i>}, q_{<i>}, p_{<i>})$, $i = 1, \cdots, m$. See (Jrgensen and Goegebeur 2002) for details.

4.4 **Advantages and disadvantages of PLS**

1. One of the advantages of PLS is the capability of predicting multiple dependent variables simultaneously. By contrast, multiple OLS regression predicts each response variable at a time. Moreover, it overcomes the difficulty of collinearity between the variables. Furthermore PLS can cope with small sample size.

2. On the other hand, PLS has some disadvantages such as distribution-free meaning that PLS does not use any distribution properties of Y and the difficulty to explain coefficients of the regression equation using factors such as Equation (4.1).

In conclusion, the PLS regression can be considered a very good method in the prediction reason and not for interpretive reason.

4.5 **Conclusion**

The PLS model is acquired using as many factors as possible (latent variables) which can be examined methodically by the accumulative of variance table to decide the number of factors necessary for a model. Several software utilizes PLS. In our study, to predicte the parametrs SAS software (PROC PLS) is utilized, which accommodates various options. Different study (Garthwaite 1994) including our experience utilizes real data or a simulation that demonstrates a powerful way of building predictive model, particularly, in the situation when the ordinary least squares performs ineffectively or inadequately. Application of the PLS method will be illustrated in chapter 5.

5

Chapter 5

Obtaining The Actual Cotton Fiber Length By Number Using PLS

5.1 Introduction

As mentioned earlier, fiber length is considered one of the most important properties of cotton for marketing purposes and yarn processing. Most researches in this field focus on developing efficient methods to measure the length of cotton. However, nearly all efforts were concentrated on the distribution of the projecting fiber lengths and various quality parameters of cotton fiber length, such as the upper half mean length, span length, short fiber content, etc. Studies on the distribution of actual fiber lengths have not been done successfully, due to technical difficulties. Since we used "original" fiber length previously, we will use both "actual" and "original" for actual fiber length in this chapter.

As we have seen that when HVI is used to analyze a projected portion of a cotton fiber beard, the clamp used in HVI testing causes the unselected portion of the beard not to be measured. Because of this, we need to find a way to approximately predict the original real length, which is the projected length plus the unselected portion.

The main objective of this chapter is to find a suitable model for such prediction. We have tried the classical OLS regression for this data but it did not fit well. Statistically, OLS regression cannot be used for such prediction because of small sample size and multicollinearity among variables. It is known that PLS regression technique is especially useful, to certain

extent, in cases where there is a small sample size, multicollinearity among variables, or lack of normal distribution assumption. In addition, PLS approach leads to stable, accurate, and highly predictable models, even for correlated variables. In this chapter, we are going to find a model using PLS regression that can overcome the problem of the unobservable hidden part when measuring cotton length, and predict the distribution of actual fiber lengths. To this end, the same eight cotton samples are used. The proposed method in this chapter can be used for centimeters as well, by converting raw data from inches to centimeters and then applying the method. There are two samples for each kind of cotton: actual and projecting. In our earlier chapters, we used Gauss-Newton algorithm and nonparametric least squares principle to find the cotton length distributions that match the data well. Specifically, it was established that a mixture of two two-parameter Weibull distributions fits the data accurately. That is, the density function of each sample (actual or projecting) is given by

$$f(x) = \alpha f_1(x, \alpha, \lambda_1, \theta_1) + (1 - \alpha) f_2(x, \alpha, \lambda_2, \theta_2)$$

where $0 < a < 1$ and $f(x, \lambda_i, \theta_i) = \lambda_i \theta_i x^{\lambda_i - 1} e^{-\theta_i x^{\lambda_i}}, x > 0, \lambda_i > 0, \theta_i > 0$, is the probability density function of a Weibull distribution, $i = 1, 2$. Thus, the distribution of fiber length is completely determined by its five parameters. Therefore, finding the distribution of fiber length is equivalent to determining the five parameters. This chapter develops a PLS regression model that estimates the five parameters of the distribution of actual length. This new approach shows that the proposed model works efficiently.

5.2 Application on cotton fiber length by number

5.2.1 Estimation of original fiber length by PLS

The objective of the PLS model is to predict the dependent variables. To achieve that, we need to follow these steps:

- The prospective sample has to be drawn from the population.
- Use existing data to find the mixure distribution of fibers and then factors.
- These factors are used in the study to build a model.
- The factors are applied to predict the dependent variables.
- Finally, inferences are drawn from the sample to the population.

The mixture distribution contains five parameters. Table 5.1 and Table 5.2 display the parameters of both fibers. These data are used to fit PLS regression. The first row of Table 5.1 and Table 5.2 contain respectively the five parameters of the mixed Weibull distribution of cotton sample Projecting-30 (observed fiber length) and Original-30 (actual fiber length), where the latter is longer. Also the second row of Table 5.1 is the parameters for Projecting-31, and that of Table 5.2 is for Original-31, and so on. Thus, the first row of both Table 5.1 and Table 5.2 constitute a pair of observations, denoted (X_1, Y_1); second rows of Table 5.1 and Table 5.2 constitute another pair of observation, denoted (X_2, Y_2), and so on. In total we have eight pair of observations. Similar to OLS regression, our goal is to estimate Y based on the observed X. In other words, we need to estimate the actual length distribution parameters $(\alpha, \lambda_1, \theta_1, \lambda_2, \theta_2)$ from observed projecting length distribution parameters $(\alpha', \lambda_1', \theta_1', \lambda_2', \theta_2')$ using PLS regression.

Table 5.3 lists the amount of variation accounted for both individual and cumulative factors. Formulas for obtaining this table are given in next section. If six factors are used, the variation accounted 100% and 97.8% for the independent and the dependent variables, respectively.

ID	Type	α	λ_1	θ_1	λ_2	θ_2
30	Projecting	0.0438	2.0032	7.7581	2.2104	3.8975
31	Projecting	0.0544	2.9038	52.543	2.4082	3.7077
33	Projecting	0.0497	4.1522	2.5435	2.1963	3.0508
34	Projecting	0.0434	6.3001	5.8474	2.1941	2.8298
35	Projecting	0.0639	2.7945	31.340	2.3607	2.6709
36	Projecting	0.0492	5.0431	1.8705	2.1533	2.7233
37	Projecting	0.0604	2.5657	20.7361	2.1650	2.3597
38	Projecting	0.0731	2.1792	2.2907	5.0285	1.4940

Table 5.1: Mixed Distribution Parameters from Projecting Length

Table 5.1 consists of 8 samples of projecting fiber length by number that contains the five parameters α, λ_1, θ_1, λ_2, θ_2. These parameters are the basic component of the mixture of Weibull distribution.

ID	Type	α'	λ_1'	θ_1'	λ_2'	θ_2'
30	Original	0.5600	1.8437	2.5948	4.3892	1.7057
31	Original	0.4236	1.6834	2.4189	4.3500	1.7956
33	Original	0.3009	1.8814	2.2543	4.0418	1.1020
34	Original	0.4917	1.8414	1.8801	5.1776	0.7985
35	Original	0.4852	1.9211	1.9902	4.9314	0.7200
36	Original	0.4578	1.7431	1.9825	5.0958	0.5340
37	Original	0.3770	1.7375	2.4820	4.6319	0.5486
38	Original	0.4052	1.6817	2.0522	5.3040	0.3853

Table 5.2: Mixed Distribution Parameters from Original Length

Table 5.2 consists of 8 samples of original fiber length by number that contains the five parameters α', λ_1', θ_1', λ_2', θ_2'.

Percent Variation Accounted for by Partial Least Squares Factors				
Number of extracted	Model Effects		Dependent Variables	
Factors	Current	Total	Current	Total
1	96.4471	96.4471	46.6810	46.6810
2	3.1414	99.5885	47.2882	93.9692
3	0.2897	99.8782	1.2136	95.1828
4	0.1205	99.9987	1.2241	96.4069
5	0.0013	100.0000	0.7934	97.2003
6	0.0000	100.0000	0.6014	97.8017

Table 5.3: Variance of X and Y Explained by the Factors

5.2.2 **Selection of the number of factors**

We now use the proportion of the total variance accounted by the model to determine the number of factors needed. Using the same notations as in PLS algorithm in Chapter 4, the above table is constructed by the following method.

The variance in X accounted for by factor i is

$$\frac{t_{<i>}^T t_{<i>} p_{<i>}^T p_{<i>}}{tr(X^T X)}. \tag{5.1}$$

The cumulative variance in X accounted for by the model with p factors is hence

$$\frac{\sum_{i=1}^{p} t_{<i>}^T t_i p_{<i>}^T p_{<i>}}{tr(X^T X)} \tag{5.2}$$

The Y-variance accounted for by the model with p factors is

$$1 - \frac{tr((Y - \hat{Y})^T (Y - \hat{Y}))}{tr(Y^T Y)} \tag{5.3}$$

The number of Factors p should be chosen such that the percentage variation explained is large enough for both X and Y. For our study, we found that five factors fit the data best.

For our data the PLS regression equation with five factors is

$$\hat{Y} = ZB \tag{5.4}$$

where $Z = \begin{pmatrix} 1 \\ X \end{pmatrix}^T$

where matrix B is:

$$B=\begin{pmatrix} 1.242455 & -0.14322 & -20.75486 & 8.815067 & -1.08788 \\ -11.48905 & 24.76514 & 274.27936 & -47.0975 & -11.4004 \\ -0.028348 & 0.053074 & 0.53171 & 0.01363 & -0.04426 \\ 0.002026 & -0.00728 & -0.0852 & 0.01329 & 0.00593 \\ 0.03255 & -0.09499 & -0.6346 & 0.2183 & 0.26097 \\ -0.067848 & 0.26520 & 3.32212 & -0.816 & 0.72028 \end{pmatrix}$$

For comparison purpose, we applied the PLS regression, equation (5.4), for all eight cotton samples. Table 5.4 shows the estimated parameters. The PDF graphs from the data Table 5.2 and from the PLS prediction Table 5.4 are presented below. A quick comparison between Table 5.2 and Table 5.4 shows that the proposed PLS approach gives reasonable estimates.

ID	$\hat{\alpha}$	$\hat{\lambda_1}$	$\hat{\theta_1}$	$\hat{\lambda_2}$	$\hat{\theta_2}$
30	0.50088	1.89203	3.54949	4.26797	1.68321
31	0.48904	1.95103	2.43263	4.97833	1.53082
33	0.44815	1.81519	2.94534	4.62569	0.99319
34	0.45037	1.92901	2.66745	5.30796	0.62853
35	0.39763	1.62006	2.04437	4.32946	0.99674
36	0.42627	1.78267	2.67496	4.76537	0.72304
37	0.36095	1.46661	1.99217	3.89528	0.86412
38	0.41552	1.76771	2.18543	5.49448	0.29277

Table 5.4: Estimation of Distribution Parameters by Number Using PLS

5.3 **Estimation of some quality parameters by number using PLS**

Let μ_i $(i = 1, 2)$ denote the mean of each Weibull distribution, and the mean of the mixture of these two Weibull distributions determined by Equation (3.3) in Chapter 2. The UHML is the mean length of those fibers longer than the median value of the weight distribution

of fibers, which is defined by equation (3.4). The lower half mean length (LHML) is defined by equation (3.5). The short fiber content (SFC) by number is given by equation (3.7).

Plugging the predicted distribution by number using PLS into the above mentioned formulas will result into the estimated mean length, UHML, LHML, SFC, CV, and UI, which in turn are given in Table 5.5. In this table, there are three rows of values. The top row is from the data, the second row is from the fitted distribution, and the third row is from the predicted distribution using the PLS model. We can see that the parameters obtained from the raw data, the mixed Weibull distribution, and the PLS model are very close.

ID 30	Mean	UHML	LHML	CV (%)	SFC (%)	UI (%)
Experimental	.658	.983	.494	45.4	31.7	80.8
Weibull Distributionn	.652	.979	.488	.455	.323	.803
PLS	.654	.977	.492	.449	.315	.805
ID 31						
Experimental	.691	.997	.528	.423	.261	.817
Weibull Distributionn	.683	.993	.520	.430	.273	.815
PLS	.682	.996	.518	.434	.278	.814
ID 33						
Experimental	.737	1.105	.553	.458	.267	.807
Weibull Distributionn	.792	1.108	.619	.385	.183	.821
PLS	.787	1.122	.607	.408	.205	.818
ID 34						
Experimental	.804	1.153	.617	.419	.208	.819
Weibull Distributionn	.798	1.148	.612	.420	.212	.818
PLS	.803	1.154	.615	.421	.211	.819
ID 35						
Experimental	.810	1.168	.620	.419	.205	.815
Weibull Distributionn	.806	1.163	.616	.419	.210	.815
PLS	.818	1.163	.630	.405	.194	.819
ID 36						
Experimental	.845	1.233	.642	.436	.208	.815
Weibull Distributionn	.839	1.228	.637	.439	.213	.815
PLS	.834	1.192	.641	.411	.194	.818
ID 37						
Experimental	.857	1.257	.650	.441	.205	.814
Weibull Distributionn	.847	1.256	.655	.441	.212	.806
PLS	.837	1.248	.629	.454	.228	.809
ID 38						
Experimental	.899	1.299	.687	.432	.192	.821
Weibull Distributionn	.892	1.293	.681	.435	.197	.820
PLS	.891	1.293	.680	.436	.198	.820

Table 5.5: Estimation of Some Length Quality Parameters by Number
It is seen that the parameters obtained from the three methods are very close.

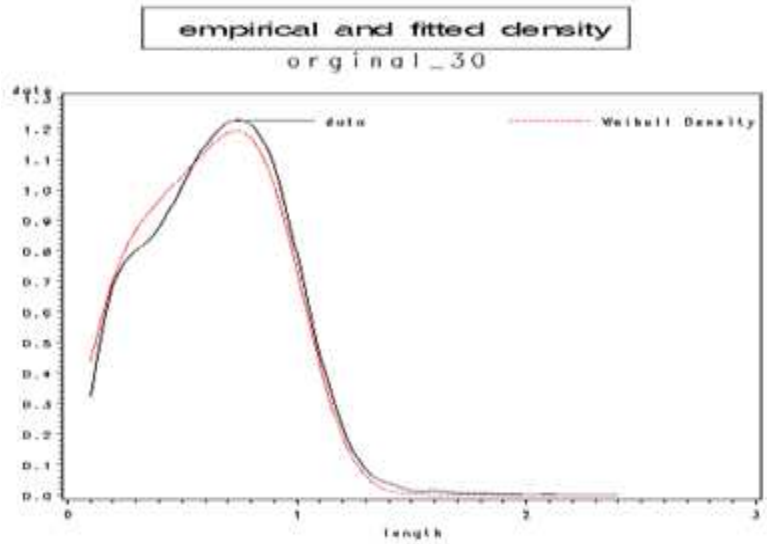

Figure 5.1: Probability density functions by number of ID 30 original fibers.

Graphical comparisons between the PDF of the estimated mixture distributions using PLS and the PDF of data are performed for original-30 fiber length (by number). One can see that mixture distribution by PLS regression provides a good match with the actual data fiber length by number.

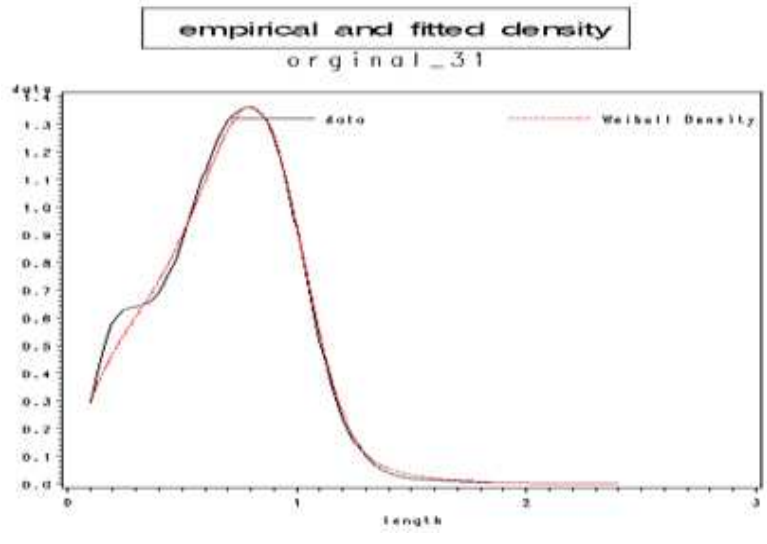

Figure 5.2: Probability density functions by number of ID 31 original fibers.

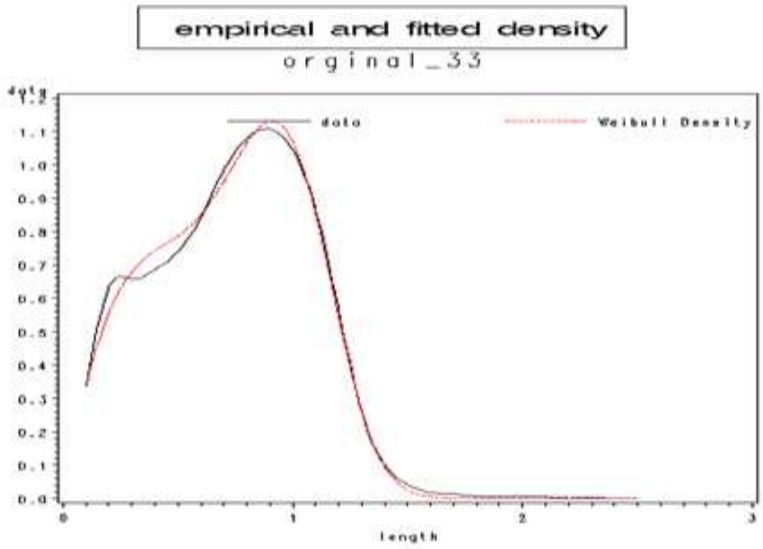

Figure 5.3: Probability density functions by number of ID 33 original fibers.

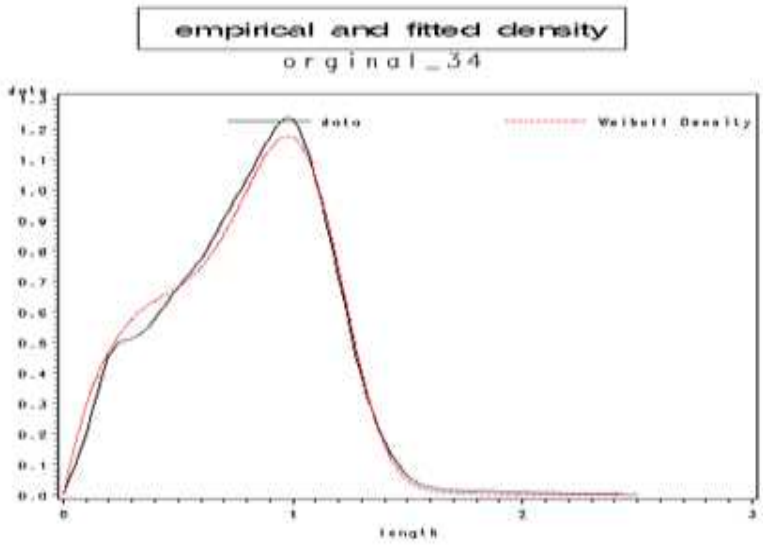

Figure 5.4: Probability density functions by number of ID 34 original fibers.

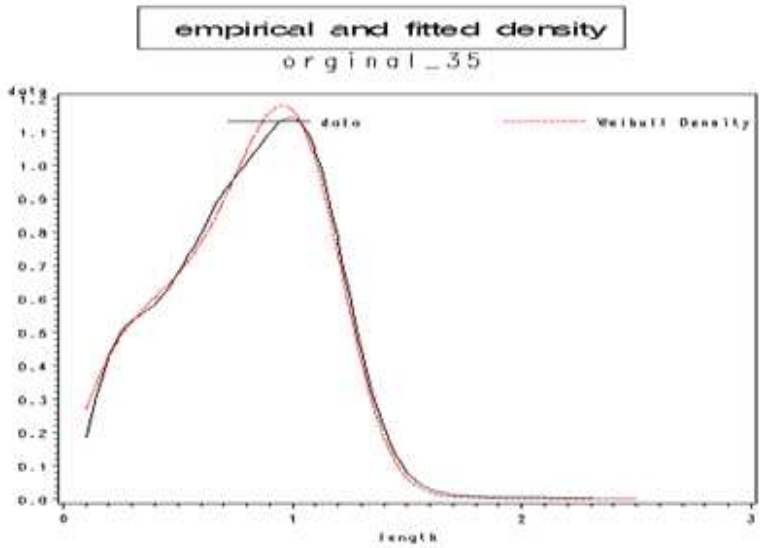

Figure 5.5: Probability density functions by number of ID 35 original fibers.

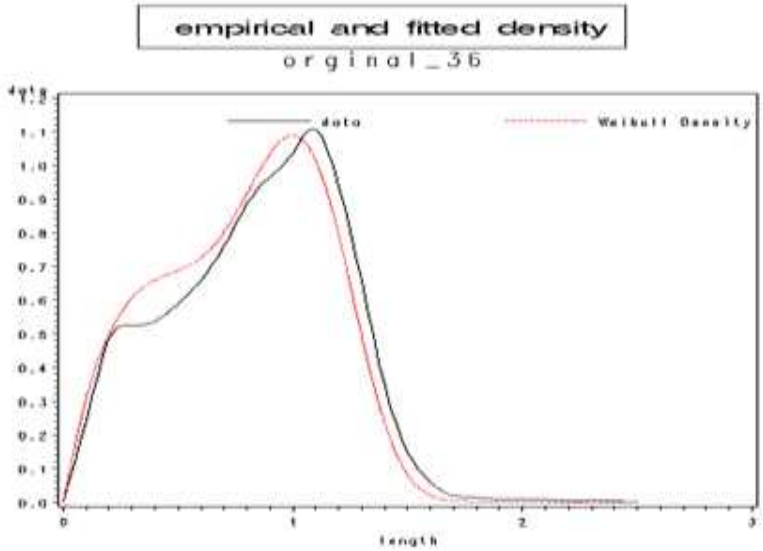

Figure 5.6: Probability density functions by number of ID 36 original fibers.

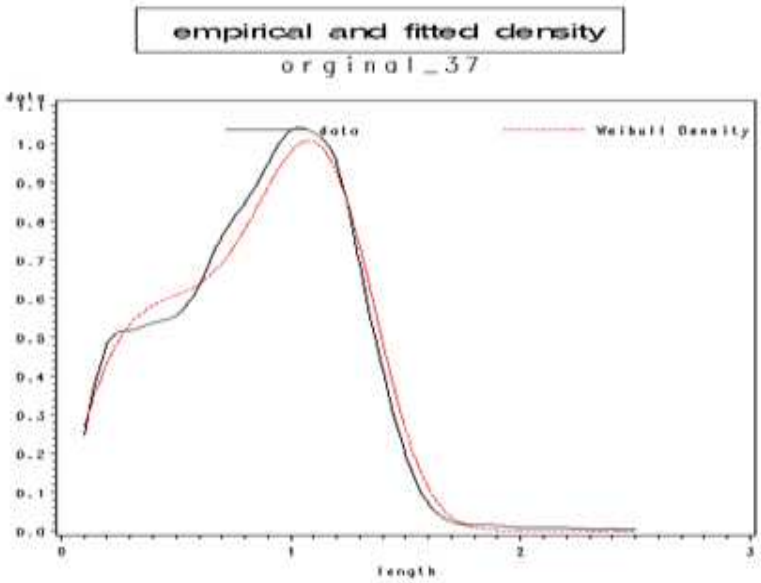

Figure 5.7: Probability density functions by number of ID 37 original fibers.

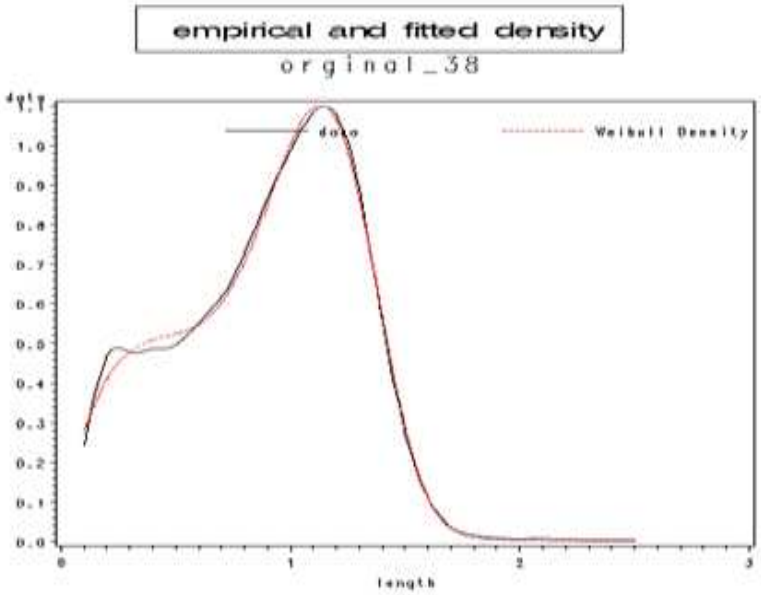

Figure 5.8: Probability density functions by number of ID 38 original fibers.

5.4 Conclusion

On one hand, as shown in Chapter 2, the distribution of cotton fiber lengths can be modeled precisely by a mixture of two two-parameter Weibull distributions. On the other hand, the present chapter provides an approach to estimating the distribution of the actual fiber length based on observed projected length. The actual or original fiber length cannot be obtained in practice, and only projected length can. Meanwhile, knowing the distribution of original fiber length is critical in the cotton industry due to obvious reasons. Therefore, finding the distribution of original fiber lengths becomes a necessity.

The method proposed in this chapter is partial least squares regression, where distribution parameters of projecting length are independent variables and that of original length are dependent variables. Graphs of density functions obtained from the proposed PLS regression method show a close match with graphs of empirical densities obtained from the experimental data. Comparing some commonly used quality parameters from Table 5.5 (experimental data, mixture distribution, and PLS regression), provide a good support of the proposed method (PLS).

We believe that this study is just a beginning. There are more studies remained to be done. For example, our current study is based on eight samples. In order to cover a wider range of cottons, studies with more samples are needed to verify and possibly modify the method proposed in this chapter. In addition, currently we are using the mixed Weibull distribution to fit data which contains 5 parameters. It is essential if a distribution with less number of parameters (< 5) can be found to fit the data, because the less the parameters the higher the precision PLS regression may provide. As far as we know, the proposed approach is new. In short, this chapter provides a method to find the distribution of actual cotton fiber lengths.

6

Chapter 6

Distribution of Cotton Fiber Length By Weight

6.1 Introduction

As stated in the previous chapter, the methodology can be used for fiber length by weight. Cui and Calamari (1998) mention that assuming the fiber linear density and its length are statistically independent, one can calculate the probability density function of fiber length by number from that by weight and visa versa. Also the length by weight distribution can be calculated from the fibrogram by taking the ratio of the area between the fibrogram curve and tangent and the total area under the fibrogram curve as shown in Figure 1.2. The mean length by weight can be found from the mean length by number and the standard deviation (Chu and Riley 1997) :

$$Mean\ Lenght\ by\ Weight\ =\ ML + \frac{(SD)^2}{ML}$$

where ML is the mean length by number, and SD is the standard deviation of the length distribution by number.

6.2 Application on cotton fiber length by weight

6.2.1 Data gathering procedure

Following the same steps as in Chapter 5, the distribution of projecting fiber length by weight and actual fiber length by weight can be modeled by a mixture of two two-parameter Weibull distributions. This mixed distribution contains five parameters. In the first part of the study in Chapter 2, we obtained parameters for both projecting and actual fiber lengths of all eight cotton samples. As shown below, the parameters are displayed by inches in Table 6.1 and Table 6.2, which are used in PLS regression. As presented previously for fiber length by number, the first rows of both Tables of the fiber length by weight will form a pair of observation, denoted by (X_1, Y_1), second rows denoted by (X_2, Y_2), and so on. In total, we have eight pairs of observation. The purpose is to estimate actual length distribution parameters $(a, \lambda_1, \theta_1, \lambda_2, \theta_2)$ from observed projecting length distribution parameters $(a', \lambda_1', \theta_1', \lambda_2', \theta_2')$.

Table 6.3 below lists the amount of variation accounted for each of these factors, both individual and cumulative. Using all the 6 factors, the variation accounted 100% and 98.49% for the independent and dependent variables, respectively.

ID	Type	α	λ_1	θ_1	λ_2	θ_2
30	Projecting	0.622	2.777	4.293	2.898	2.236
31	Projecting	0.986	2.885	3.142	6.126	0.267
33	Projecting	0.264	2.944	4.077	2.935	0.267
34	Projecting	0.925	2.764	2.112	4.853	5.232
35	Projecting	0.915	2.801	2.199	2.899	0.985
36	Projecting	0.935	2.788	2.079	2.872	0.864
37	Projecting	0.919	2.774	1.888	2.888	0.921
38	Projecting	0.993	1.888	1.593	17.533	1.998

Table 6.1: Mixed Distribution Parameters from Projecting Length by Weight

Table 6.1 consists of 8 samples of projecting fiber length by weight that contains the five parameters α, λ_1, θ_1, λ_2, θ_2. These parameters are the basic component of the mixture of Weibull distribution.

ID	Type	α'	λ'_1	θ'_1	λ'_2	θ'_2
30	Original	0.480	2.399	1.636	4.799	1.492
31	Original	0.305	2.153	1.397	4.784	1.532
33	Original	0.403	2.297	1.402	5.260	0.774
34	Original	0.400	2.395	1.112	5.644	0.647
35	Original	0.164	2.640	3.397	4.915	0.615
36	Original	0.439	2.484	1.010	6.098	0.366
37	Original	0.382	2.350	1.026	5.155	0.369
38	Original	0.347	2.308	0.975	6.155	0.264

Table 6.2: Mixed Distribution Parameters from Original Length by Weight

Table 6.2 consists of 8 samples of original fiber length by number that contains the five parameters α', λ'_1, θ'_1, λ'_2, θ'_2.

Percent Variation Accounted for by PLS Factors				
Number of extracted	Model Effects		Dependent Variables	
Factors	Current	Total	Current	Total
1	85.4675	85.4675	78.3502	78.3502
2	11.4934	96.9610	16.3538	94.7040
3	2.0264	98.9874	1.7323	96.4363
4	0.9880	99.9754	1.5874	98.0237
5	0.0244	99.9998	0.3563	98.3800
6	0.0002	100.000	0.1122	98.4922

Table 6.3: Variance of X and Y Explained by the Factors

6.2.2 **Selection of the number of factors**

Since our data in both Tables 6.1 and 6.2 has five variables, we have chosen five factors out of six. The variation summary shows that over 99.99% of the predictor variation and 98% of the response variations are accounted. The number of factors can also be determined by considering the proportion of the total variance accounted by the model. In one hand, The cumulative proportion of the total X-variance accounted for by the model with p factors is given by Equation (5.1), on the other hand, the Y-variance accounted for by the model with p components is given by Equation (5.3). The number of components p should be chosen

such that the percentage variation explained is large enough for both X and Y.

For our data the PLS regression equation is $\hat{Y} = ZB$

where $Z = \begin{pmatrix} 1 \\ X \end{pmatrix}^T$, and matrix B is

$$\text{B}= \begin{pmatrix} 1.242 & -0.143 & 6.910 & 8.815 & -1.088 \\ -11.489 & 24.765 & -50.167 & -47.098 & -11.400 \\ -0.028 & 0.0531 & -0.269 & 0.014 & -0.044 \\ 0.002 & -0.007 & 0.008 & 0.013 & 0.006 \\ 0.033 & -0.095 & -0.007 & 0.218 & 0.261 \\ -0.068 & 0.265 & -0.398 & -0.816 & 0.720 \end{pmatrix}$$

For comparison purpose, we applied the PLS regression equation (4.2) for all eight cotton samples and present the estimated parameters in Table 6.4. The graphs of PDF's from the data (i.e. with parameters given in Table 6.2 and from the PLS prediction, i.e. with parameters given in Table 6.4, are presented below. A quick comparison between Tables 6.2 and 6.4 shows that the proposed PLS regression gives reasonable estimates. More comments and discussions are given in the conclusion section.

ID	$\hat{\alpha}$	$\hat{\lambda_1}$	$\hat{\theta_1}$	$\hat{\lambda_2}$	$\hat{\theta_2}$
30	0.48554	2.39152	1.57939	4.81509	1.49114
31	0.28500	2.19906	1.65785	4.75767	1.52040
33	0.39542	2.30429	1.47374	5.23375	0.77695
34	0.39206	2.40757	1.20204	5.62465	0.64623
35	0.33081	2.44418	1.70473	5.46135	0.56823
36	0.33419	2.46908	1.72106	5.52150	0.49510
37	0.34472	2.51262	1.70672	5.68755	0.29307
38	0.35185	2.29568	0.90677	6.16283	0.26618

Table 6.4: Estimation of Parameters Distribution by Weight Using PLS

6.3 Estimation of some quality parameters of fiber length by weight

Plugging the predicted distribution obtained from PLS into these relevant formulas, the estimated mean length, SFC, and CV are obtained and are given in Table 6.5. For each fiber sample in Table 6.5 there are three rows of values: The top row is from the data, the second row is from the fitted mixture distribution, and the third row is from the predicted distribution using PLS. From Table 6.5 we see that the parameters obtained from the raw data and that obtained from the mixed Weibull distribution, and that obtained from the PLS model are very close. The graphs of the PDF's by weight for Sample ID 30 to 38 are shown in Figure 6.1 to Figure 6.8. The PDF's by weight using actual fiber length as well as the PDF's from the PLS prediction are presented side to side. It can be seen that both curves between the sample and the estimated are in good agreement.

ID 30	Mean	CV (%)	SFC (%)
Data	.791	.357	15.3
Mixed Weibull	.785	.349	15.5
PLS	.798	.3476	15.30
ID 31			
Data	.814	.329	11.82
Mixed Weibull	.813	.328	12.0
PLS	.800	.319	12.54
ID 33			
Experimental	.892	.343	10.95
Weibull Distribution	.885	.333	11.2
PLS	.880	.332	11.43
ID 34			
Experimental	.945	.323	8.26
Weibull Distributionn	.935	.315	8.39
PLS	.929	.314	8.74
ID 35			
Experimental	.952	.322	8.13
Weibull Distributionn	.938	.307	8.58
PLS	.921	.315	9.76
ID 36			
Experimental	1.005	.321	7.56
Weibull Distributionn	1.002	.314	7.55
PLS	.936	.319	9.64
ID 37			
Experimental	1.024	.326	7.30
Weibull Distributionn	1.017	.318	7.43
PLS	1.023	.331	9.29
ID 38			
Experimental	1.067	.308	6.42
Weibull Distributionn	1.064	.305	6.40
PLS	1.072	.305	6.18

Table 6.5: Estimation of Some Length Quality Parameters by Weight

From the table we see that the parameters obtained from the three methods are very close. The difference is at the second decimal place for all parameters.

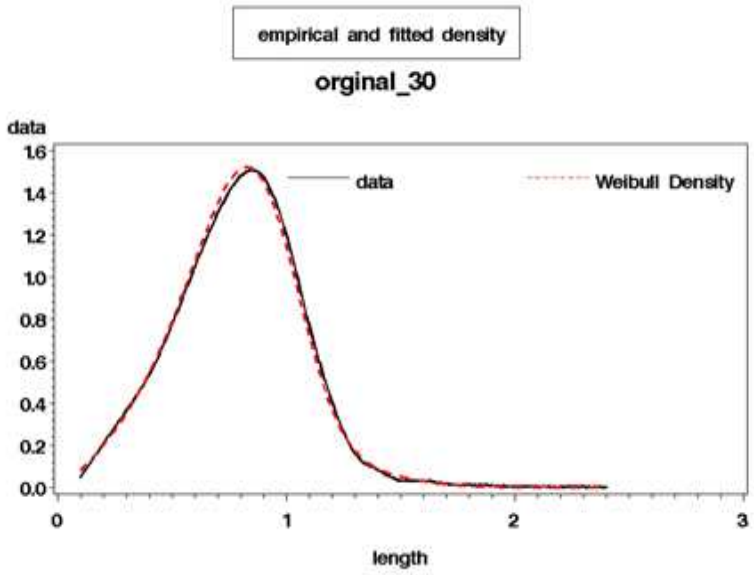

Figure 6.1: Probability density functions by weight of ID 30 actual fibers.

It is seen that mixture distribution obtained by PLS regression provides a good match with the actual data fiber length by weight.

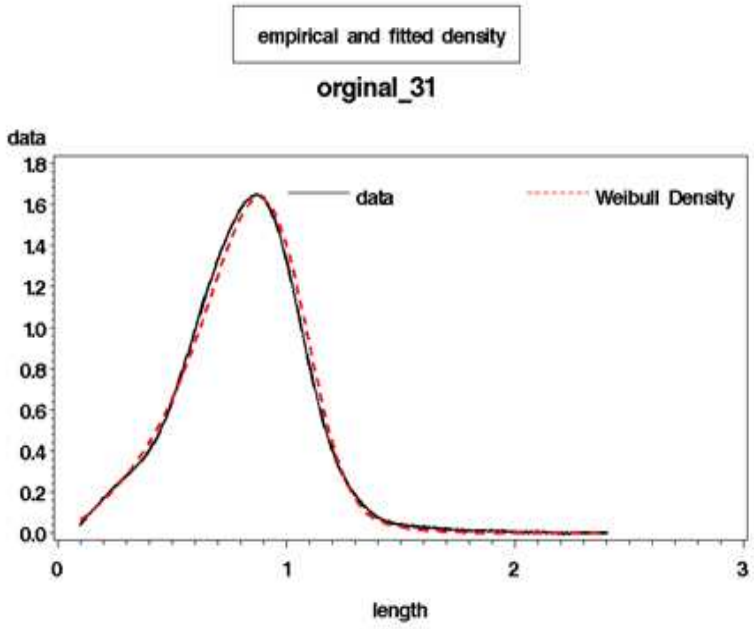

Figure 6.2: Probability density functions by weight of ID 31 actual fibers.

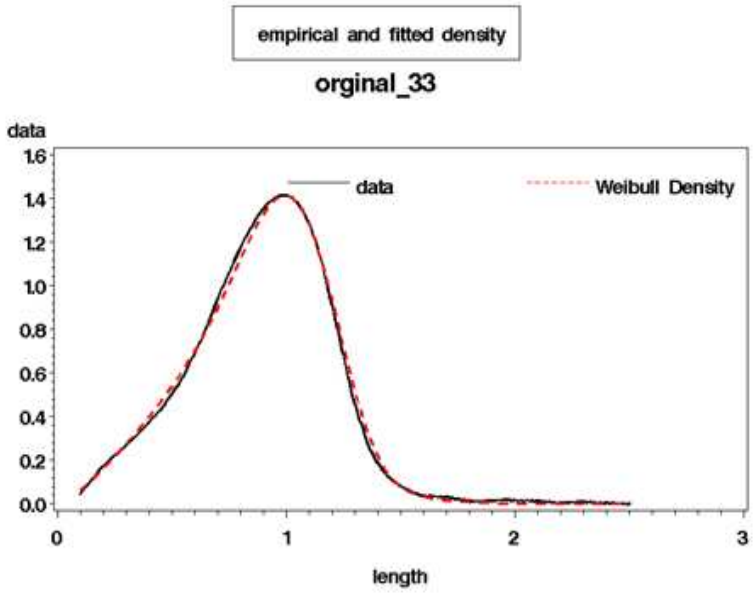

Figure 6.3: Probability density functions by weight of ID 33 actual fibers.

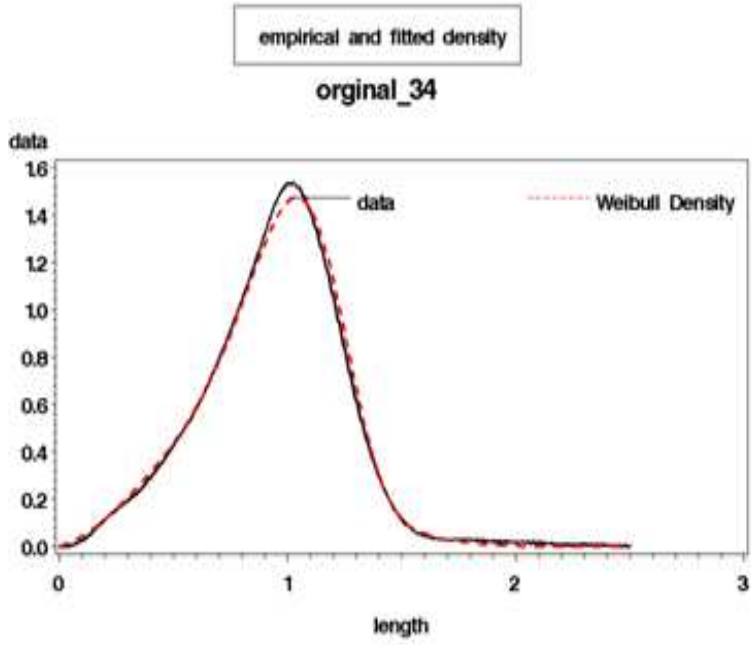

Figure 6.4: Probability density functions by weight of ID 34 actual fibers.

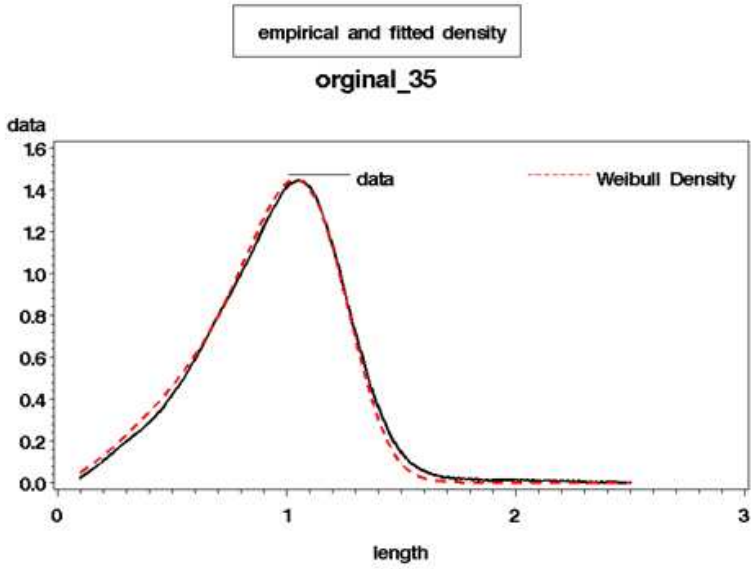

Figure 6.5: Probability density functions by weight of ID 35 actual fibers.

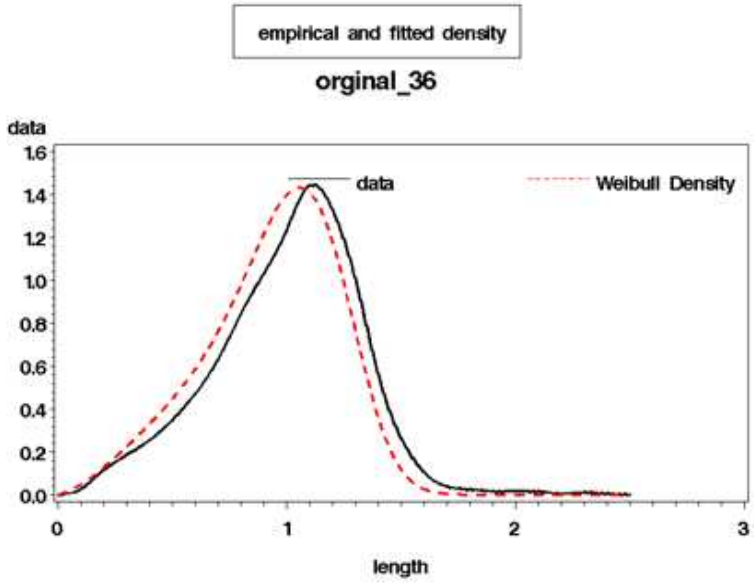

Figure 6.6: Probability density functions by weight of ID 36 actual fibers.

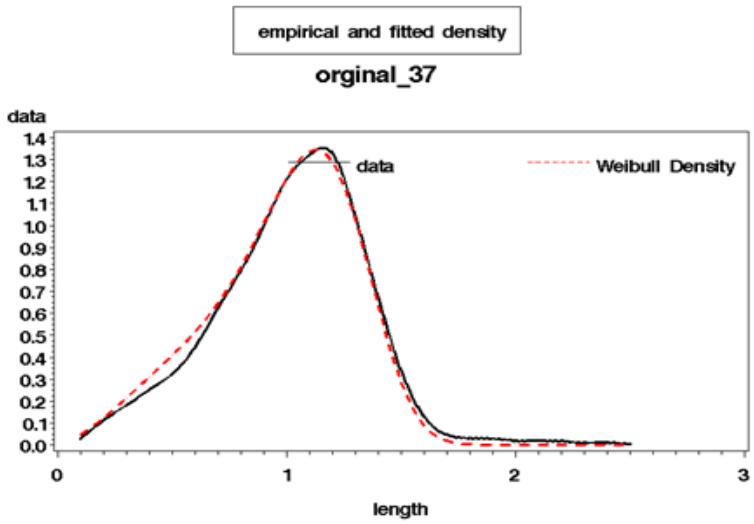

Figure 6.7: Probability density functions by weight of ID 37 actual fibers.

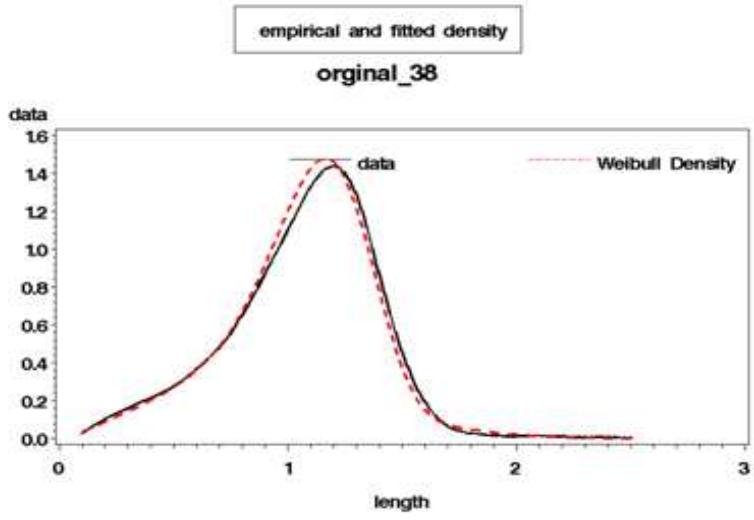

Figure 6.8: Probability density functions by weight of ID 38 actual fibers.

6.4 Conclusion

In this chapter, it was shown that the distribution of cotton fiber lengths by weight can be modeled precisely by a mixture of two two-parameter Weibull distributions. The present chapter provides an approach to estimating the distribution of the actual fiber length by-weight based on observed projecting lengths. As mentioned earlier, actual fiber length cannot be obtained in practice, therefore, knowing the distribution of actual fiber length is critical in the cotton industry.

Partial least squares regression is the method that was used in this chapter, where distribution parameters of projecting lengths are independent variables and those of original length are dependent variables. Graphs of density functions obtained from the proposed PLS regression showed a close match with graphs of empirical densities obtained from the raw data. Comparisons of some commonly used quality parameters Table 6.5 among data, mixture distribution and PLS regression provide a good support of the proposed method as well. In the author's opinion, this study is an exploratory study. More studies remain to be done.

7

Chapter 7

A New Model For The Distribution of Fiber Length By Weight

7.1 Introduction

As mentioned earlier the relation between length by number and length by weight is defined by $h_w(x) = \frac{xh(x)}{\overline{x}}$, where $h(x)$ denotes the probability density function of fiber length by number, that is, $h(x)$ is a frequency function of lengths, and $\overline{x}$ denotes the sample mean of a data set. The frequency function by number and that by weight are equivalent, i.e., knowing one implies knowing the other. It is known that reducing the number of parameters in a regression can enhance the estimation of parameters. That is why we are considering introducing a new distribution with only three parameters to describe the distribution of fiber lengths by weight. It is found that the mixture of Weibull distributions fits the data very well; on the other hand, one Weibull distribution does not fit. Therefore it is worth trying to generalize a Weibull distribution by adding one more parameter to make it more flexible to fit data. After tried different forms of distributions, we introduce a new distribution that has three parameters. We found that the new distribution fits the data very well. Unfortunately, the new distribution works only for length by weight, not by number. The reason is that the PDF of length by weight does not have a "hunk" on the left side (shorter fiber portion) of the density curse; in contrast the PDF of length by number usually has such a hunk.

7.2 New Distribution

7.2.1 Definition and properties of the new distribution

A continuous random variable X follows (new) distribution $\text{ND}(a, b, c)$ if its PDF is given by

$$f_X(x) = abcx^{c-1}e^a e^{bx^c} e^{-ae^{bx^c}}, x \geq 0, a > 0, b > 0, c > 0 \tag{7.1}$$

The CDF of $\text{ND}(a, b, c)$ is

$$F_X(x) = P(X \leq x) = 1 - e^{a(1-e^{bx^c})} \tag{7.2}$$

It can be seen that $y = e^{x^c}$ has a Weibull distribution with parameters a and b. The expected value of X is given by

$$\mu = E(X) = \int_0^\infty [1 - F_X(x)]\, dx = \int_0^\infty e^{a(1-e^{bx^c})}\, dx$$

In order to illustrate the flexibility of the new distribution we present graphs of $\text{ND}(a, b, c)$ with various parameter values.

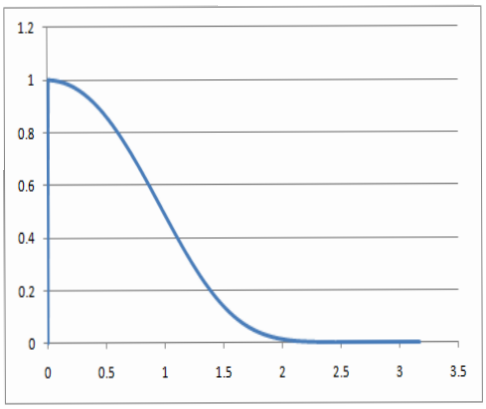

Figure 7.1: Plot of ND(1, 1, 1)

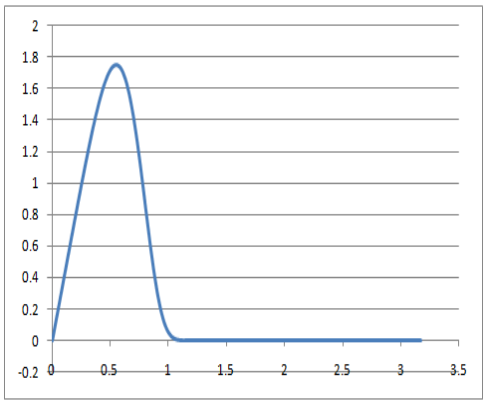

Figure 7.2: Plot of ND(1, 1, 2)

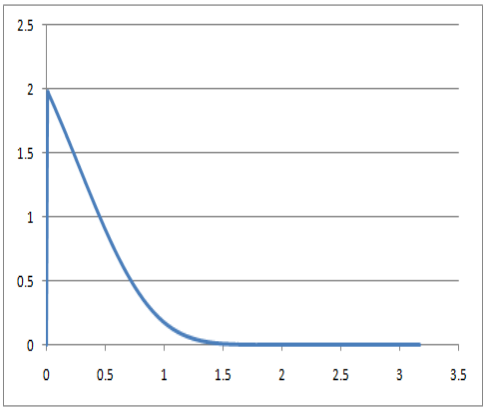

Figure 7.3: Plot of ND(2,1, 1)

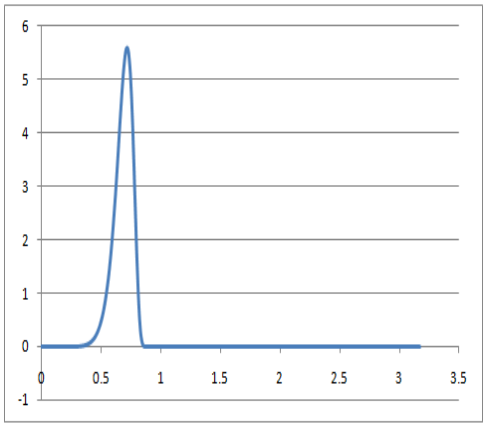

Figure 7.4: Plot of ND(5, 5, 10)

We show in this section that the second moment of X is finite, since we need variance and expected value in our calculation of estimation of some quality parameters. The second moment is given by

$$E(X^2) = \int_0^\infty abcx^2x^{c-1}e^a e^{bx^c} e^{-ae^{bx^c}} dx.$$

Let $y = e^{bx^c}$. Then $ln(y) = bx^c$ and $dy = e^{bx^c}bcx^{c-1}dx$.

$$E(X^2) = K\int_1^\infty \frac{x^{c+1}ye^{-ay}}{x^{c+1}y}dy = K\int x^2e^{-ay}dy,$$

$$E(X^2) = K\int_1^\infty ln(y)^{\frac{2}{c}}e^{-ay}dy,$$

where K is a constant.

Since $ln(y) \leq y$ for all $y > 1$, we have

$$K\int_1^\infty ln(y)^{\frac{2}{c}}e^{-ay}dy < K\int_1^\infty y^{\frac{2}{c}}e^{-ay}dy < \infty.$$

This was to be shown.

Maximum Likelihood Estimators

In this section the maximum likelihood estimation of the parameters $a,\ b,\ c$ of distribution ND(a, b, c) are considered. That is, we consider estimation of a, b, and c when all are unknown. If $x_1, ..., x_n$ is a random sample from ND(a, b, c) , then the log-likelihood function, $L(a, b, c)$, is given by

$$L(a,b,c) = n\ln(a)+n\ln(b)+n\ln(c)+(c-1)\sum_{i=1}^{n}\ln(x_i)+na+n\ln(b)+c\sum_{i=1}^{n}x_i-a\sum_{i=1}^{n}e^{bx_i^c} \quad (7.3)$$

The normal equations become:

$$\frac{\partial L}{\partial a} = \frac{n}{a} + n + \sum_{i=1}^{n} e^{bx_i^c} = 0. \tag{7.4}$$

$$\frac{\partial L}{\partial b} = 2\frac{n}{b} - ab\sum_{i=1}^{n} x_i^c e^{bx_i^c} = 0. \tag{7.5}$$

$$\frac{\partial L}{\partial c} = \frac{n}{c} + \sum_{i=1}^{n} \ln(x_i) + \sum_{i=1}^{n} x_i - abc\sum_{i=1}^{n} x_i^{(c-1)} e^{bx_i^c} = 0. \tag{7.6}$$

From (7.5), we obtain a as a function of b and c, say $\hat{a}(b, c)$, where

$$\hat{a}(b, c) = \frac{1}{n}(\sum_{i=1}^{n} e^{bx_i^c} - n). \tag{7.7}$$

Putting $\hat{a}(b, c)$ in (7.3), we obtain

$$L(\hat{a}(b,c), b, c) = n\ln(\hat{a}) + n\ln(b) + n\ln(c) + (c-1)\sum_{i=1}^{n}\ln(x_i) + n\hat{a} + n\ln(b) + c\sum_{i=1}^{n} x_i - \hat{a}\sum_{i=1}^{n} e^{bx_i^c} \tag{7.8}$$

Therefore, MLE of b and c, can be obtained by maximizing equation (7.8) with respect to b and c. It can be seen that if $X \sim \text{ND}(a, b, c)$ then when $b = 1$ $\exp(X)$ has a Weibull distribution with parameters a and c.

Random Number Generator

Random numbers of distribution $ND(a, b, c)$ can be obtained from random variable $U(0, 1)$ by using the inverse distribution. For $ND(a, b, c)$, its inverse is obtained as follows.

Given,

$$x = F(y) = \int_0^y ND(z)\, dz$$

we get

$$x = F(y) = 1 - e^{a(1-e^{bx^c})} \tag{7.9}$$

Inverting Equation 7.9 to write $y = G(x)$, we have

$$y = G(x) = \sqrt[c]{\frac{1}{b}\ln(1 + \frac{1}{a}\ln(\frac{1}{1-x}))}. \tag{7.10}$$

Therefore; equation 7.10 generates a random value $y \sim ND(a, b, c)$once a random value $x \sim U(0, 1)$ is generated.

7.3 Goodness-of-fit test

Following the same steps as in Chapter 2, non-linear least squares regression models were constructed with the new distribution. In other words, we repeated the procedure in Chapter 2 by replacing the mixture of Weibull distributions with ND(a, b, c). The results showed that our new distribution is in good agreement with the available data of fiber length by weight. Therefore, each of the eight samples can be characterized by the new distribution with appropriate parameters determined by the non-linear LS regression. We present the estimated parameters of fitted distributions for Original and Projecting lengths in Table 7.1. Kolmogorov-Smirnov goodness-of-fit test was performed to verify that the new distribution does fit the data. Let the hypothesis be "the data follows distribution ND(a, b, c)". We follow the same steps as in Chapter 2 and use sample Original ID-30 as an example to re-present them below.

1. Randomly re-sample 2500 observations from Original ID-30 data set.

2. The fitted CDF for Original ID-30 is ND(1.7384, 0.6514, 2.8169), and the corresponding ECDF of the re-sampled data is given by Equation (2.1).

3. Use $\alpha = 0.10$ and compare D_n (See Chapter 2) with $d_\alpha \simeq 1.22$. If D_n is less than 1.22, the Kolmogorov-Smirnov statistic is not significant, i.e. the hypothesis that these 2500 re-sampled data points follow distribution ND(1.7384, 0.6514, 2.8169) is accepted.

4. Repeat steps 1 to 3 500 times and record the number of times that the hypothesis is accepted.

The above steps were performed for both Original and Projecting fiber lengths of each of the eight samples. Out of 500 repetitions the percentage of acceptance is higher than 95% every time. Therefore, we can claim that the fiber length distribution by weight can be described by distribution ND(a, b, c). In addition, for each of the eight cottons, graphical comparisons between the PDF of the estimated ND(a, b, c) and the PDF of data are performed as presented in Figures 7.5 to 7.12.

ID	Type	a	b	c
30	Original	1.5	0.7	2.7
	Projecting	50.009	0.06	2.7
31	Original	0.96	0.92	2.75
	Projecting	59.68	0.048	2.825
33	Original	0.448	1.184	2.259
	Projecting	54.469	0.04	2.79
34	Original	0.835	0.699	2.87
	Projecting	53.38	0.0395	2.874
35	Original	0.778	0.705	2.74
	Projecting	54.348	0.035	2.746
36	Original	0.0175	3.6076	1.0267
	Projecting	56.82	0.0325	2.74
37	Original	0.468	0.814	2.384
	Projecting	53.51	0.031	2.716
38	Original	0.386	0.844	2.448
	Projecting	55.5	0.01	2.85

Table 7.1: Parameter Estimation of ND(a, b, c)

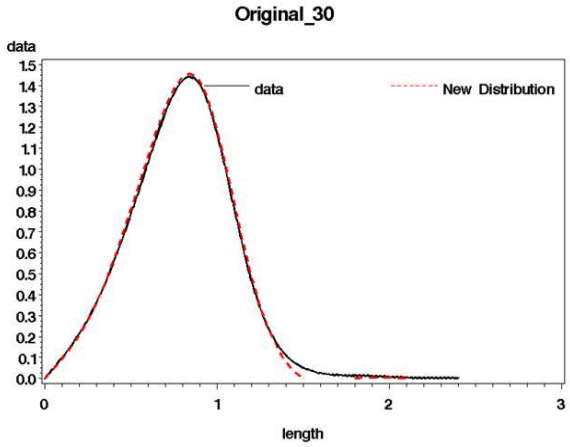

Figure 7.5: PDF by weight of original ID 30

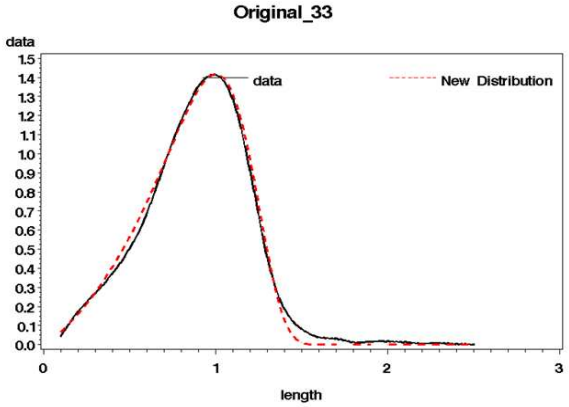

Figure 7.6: PDF by weight of original ID 33

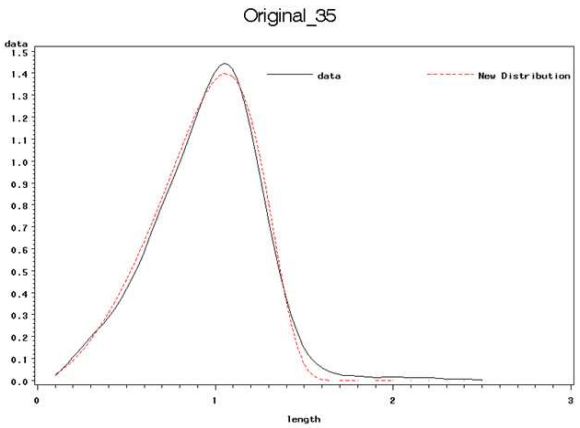

Figure 7.7: PDF by weight of original ID 35

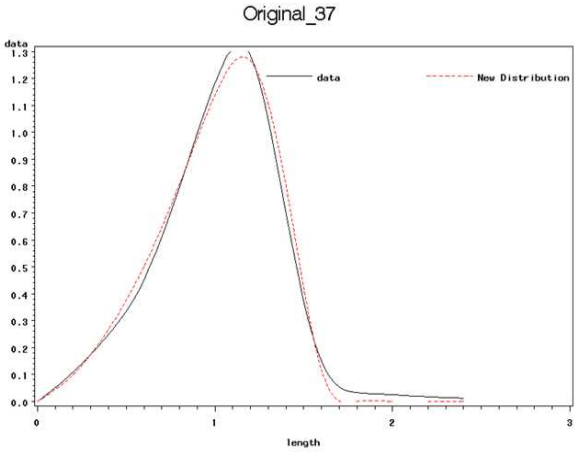

Figure 7.8: PDF by weight of original ID 37

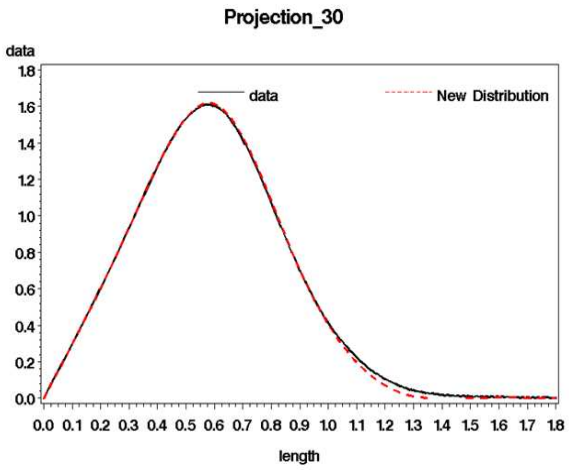

Figure 7.9: PDF by weight of projecting ID 30

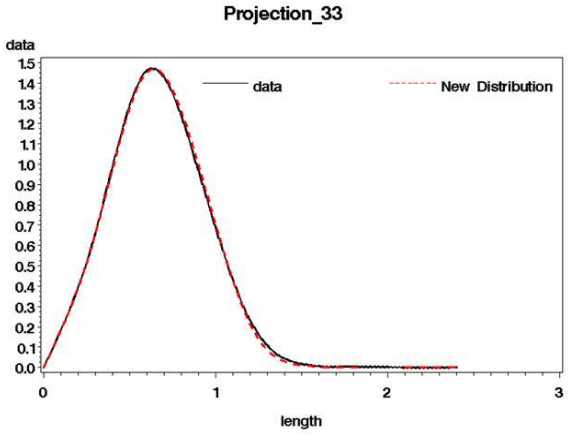

Figure 7.10: PDF by weight of projecting ID 33

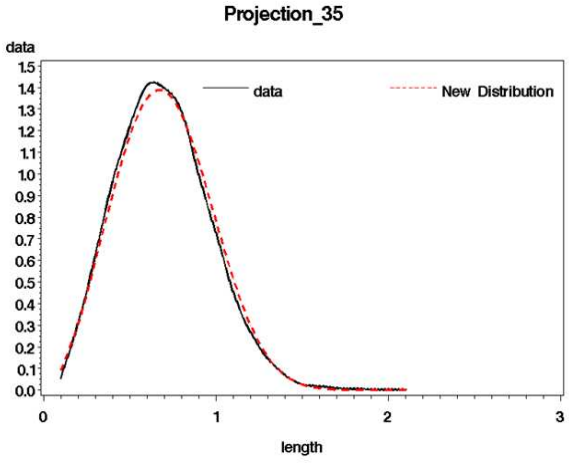

Figure 7.11: PDF by weight of projecting ID 35

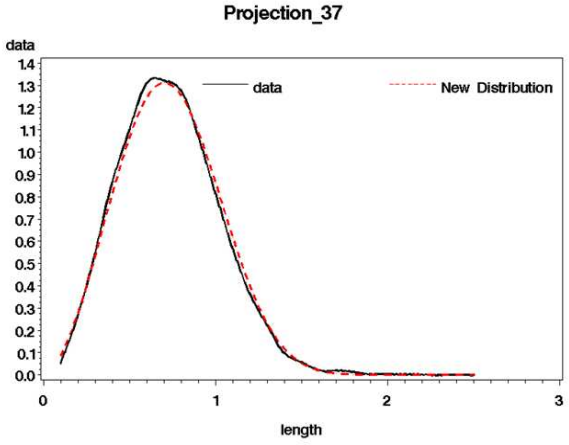

Figure 7.12: PDF by weight of projecting ID 37

7.4 **Estimation of some quality parameters**

Comparisons between the quality parameters obtained from the original data and that obtained from the estimated ND(a, b, c) are presented. The purpose of the comparison is to further investigate the fit of the estimated distribution to the data. These parameters include ML, CV, and SFC. The estimated parameters are given in Table 7.2. For each type of fiber sample, Original or Projecting, there are two rows of values in the table: the top row (Data) is from the experimental data and the second row (ND(a, b, c)) is from the fitted distribution. It is seen that parameter values obtained from the experimental data and those obtained from ND(a, b, c) are in good agreement. Considering the natural non-uniformity of cotton length, the third decimal of the data usually does not have statistical significance. If we round up the data in Table 7.2 to the second decimal, the results from data and from fitted ND(a, b, c) are almost identical, indicating an excellent fit of the new distribution.

ID 30	Mean	CV (%)	SFC (%)
Data	.791	.357	15.3
ND(a, b, c)	.781	.333	15.67
ID 31			
Data	.814	.329	11.82
ND(a, b, c)	.802	.310	13.10
ID 33			
Data	.892	.343	10.95
ND(a, b, c)	.870	.324	11.81
ID 34			
Data	.945	.323	8.26
ND(a, b, c)	.920	.300	8.04
ID 35			
Data	.952	.322	8.13
ND(a, b, c)	.932	.310	8.29
ID 36			
Data	1.005	.321	7.56
ND(a, b, c)	1.003	.318	7.58
ID 37			
Data	1.024	.326	7.30
ND(a, b, c)	1.010	.313	7.59
ID 38			
Data	1.067	.308	6.42
ND(a, b, c)	1.040	.296	6.25

Table 7.2: Estimation of Some Length Quality Parameters by Weight

7.5 Application of PLS using ND(a, b, c)

In this section, we are going to repeat procedure of Chapter 5 by replacing distribution parameters, Tables 5.1 and 5.2, of mixed Weibull with parameters of ND(a, b, c) given in Table 7.1. In particular, we will use the PLS regression to estimate parameters (a, b, c) of Original length based on the parameters of Projecting length. For our data we use 3 factors based on the variation explained by the variation of the independent and dependent variables. This variation is explained by PLS regression model based on our data that is given by Table 7.5, which lists the amount of variation accounted for each of these factors

both individual and cumulative. Note that all of the variation is accounted for by 4 factors is given by 100% for the independent variables and 99.04% for the dependent variables.

Percent Variation Accounted for by PLS Factors				
Number of extracted	Model Effects		Dependent Variables	
Factors	Current	Cumulative	Current	Cumulative
1	99.9994	99.9994	97.3893	97.3893
2	0.0005	100.0000	0.5542	97.9435
3	0.0000	100.0000	0.7666	98.7101
4	0.0000	100.0000	0.3334	99.0434

With three factors the variation summary shows that 100% of the independent variation and 98.7% of the dependent variation are accounted. The dependent variables, which are distribution parameters of Original length, can be predicted using the multiple regression formula $\hat{Y} = XB$ where $X = (1, a, b, c)$ denotes the distribution parameters of Projecting length from Table 7.1, and matrix B is

$$\text{B}=\begin{pmatrix} -0.403 & -0.735 & -1.908 \\ -0.030 & 0.019 & -0.006 \\ 21.087 & 0.615 & 8.623 \\ 0.728 & 0.172 & 1.613 \end{pmatrix}$$

PLS regression was applied for all eight cotton samples, and the estimated parameters are presented in Table 7.3. For each of the eight cottons, graphical comparisons between the PDF of the estimated using PLS and the PDF of data are performed as shown in Figures 7.13 to 7.21.

ID	λ'	θ'	β'
30	1.29812	0.73888	2.67152
31	0.84905	0.94182	2.71829
33	0.81341	0.82952	2.62283
34	0.69436	0.80312	2.63932
35	0.67965	0.81652	2.50946
36	0.54736	0.86203	2.46400
37	0.59896	0.79259	2.43140
38	0.28599	0.85034	2.51035

Table 7.3: New Distribution Parameters by PLS

7.5.1 **Estimation of some quality parameters by PLS**

Using the same procedure as in chapter 5 the quality parameters, Mean, CV and SFC, are obtained and presented in Table 7.5. We include Table 7.2 for comparison purpose. From the table we see that the quality parameters obtained from data, ND(a, b, c) and PLS are very close.

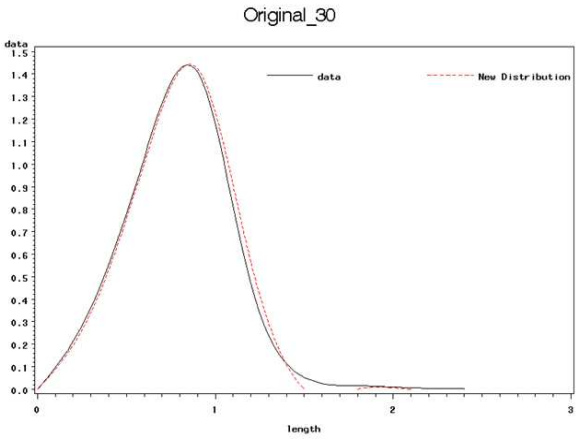

Figure 7.13: PDF by weight of original ID 30 using PLS

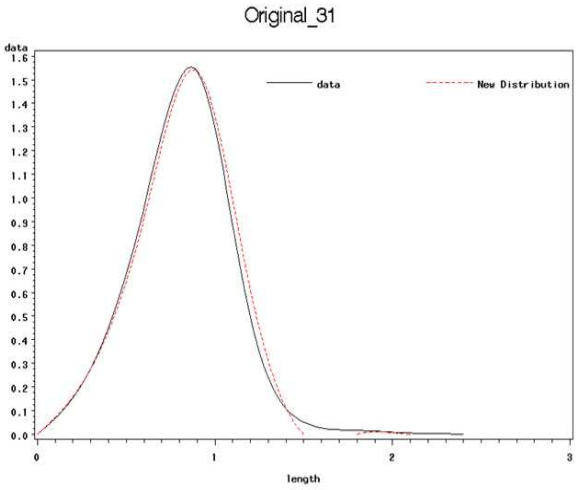

Figure 7.14: PDF by weight of original ID 31 using PLS

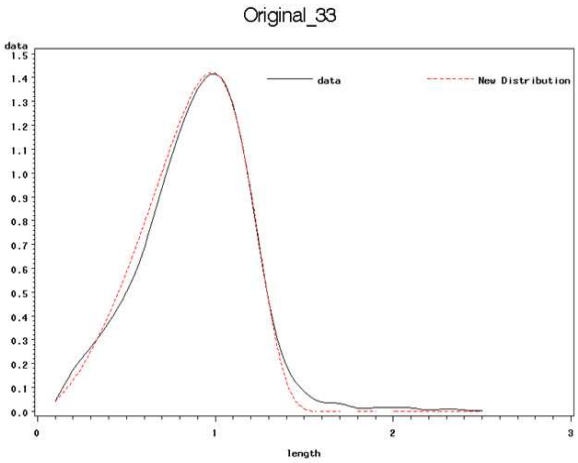

Figure 7.15: PDF by weight of original ID 33 using PLS

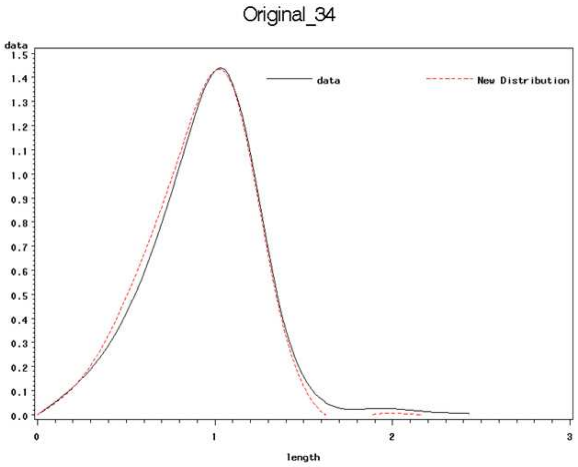

Figure 7.16: PDF by weight of original ID 34 using PLS

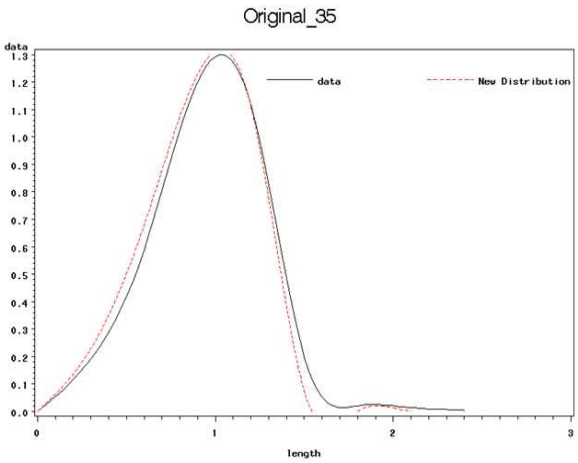

Figure 7.17: PDF by weight of original ID 35 using PLS

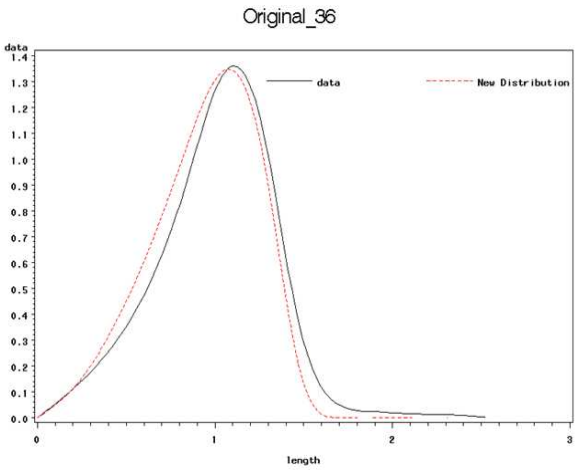

Figure 7.18: PDF by weight of original ID 36 using PLS

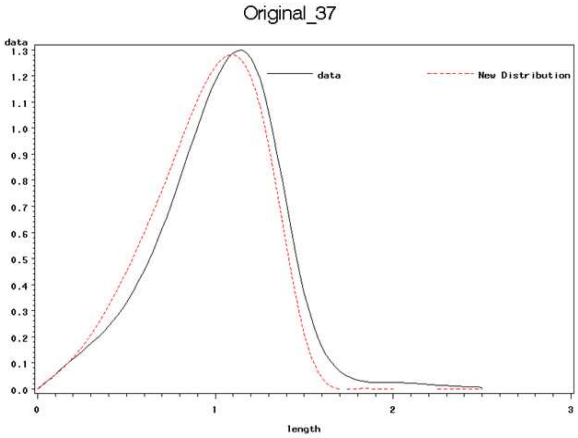

Figure 7.19: PDF by weight of original ID 37 using PLS

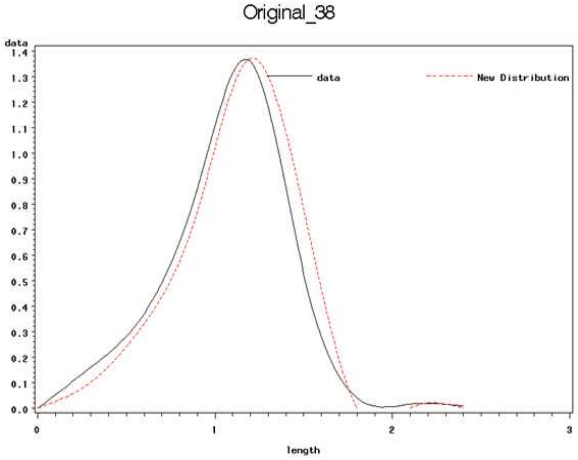

Figure 7.20: PDF by weight of original ID 38 using PLS

ID 30	Mean	CV (%)	SFC (%)
Data	.791	.357	15.3
ND(a, b, c)	.781	.333	15.67
PLS	.796	.330	14.75
ID 31			
Data	.814	.329	11.82
ND(a, b, c)	.802	.310	13.10
PLS	.819	.308	12.25
ID 33			
Data	.892	.343	10.95
ND(a, b, c)	.870	.324	11.81
PLS	.865	.315	11.06
ID 34			
Data	.945	.323	8.26
ND(a, b, c)	.920	.300	8.04
PLS	.913	.306	9.10
ID 35			
Data	.952	.322	8.13
ND(a, b, c)	.932	.310	8.29
PLS	.920	.316	9.94
ID 36			
Data	1.005	.321	7.56
ND(a, b, c)	1.003	.318	7.58
PLS	.956	.290	8.64
ID 37			
Data	1.024	.326	7.30
ND(a, b, c)	1.010	.313	7.59
PLS	.960	.320	9.19
ID 38			
Data	1.067	.308	6.42
ND(a, b, c)	1.040	.296	6.25
PLS	1.10	.280	5.50

Table 7.4: Estimation of Some Length Quality Parameters by Weight Using ND(a, b, c)

7.6 Conclusion

In this chapter, we introduced a new distribution with three parameters to model the distribution of fiber length by weight. In early chapters we used the mixture of Weibull distributions to model the distribution of fiber length by weight. Our calculations showed that the proposed new distribution works as well as the mixture of Weibull distributions. Since the new distribution has only three parameters, two parameters less than that of the mixture of Weibull distributions, the new distribution should be more efficient in PLS estimation. That is, the new distribution should work better when estimating the density of original fiber lengths by weight than the mixed Weibull distributions.

Concluding Remarks

This dissertation solves three problems.

1. **Underlying distribution of fiber length.**

   Fiber length is regarded as the most important property of cotton in marketing and yarn processing. The mathematical function describing cotton fiber length was searched both for the original sample population and the fibers picked and measured by the beard method (projecting). We found for both projecting and original fiber length that a mixture of two Weibull distributions fits the data very well. A non-parametric goodness-of-fit test also confirms the result. Therefore, this distribution can be used to characterizes the entire fiber length and all length parameters such as the average fiber length, short fiber cotton, upper-half mean fiber length and so on. Furthermore, numerical comparisons for various parameters between the distribution of length from the data and the fitted distribution show a very good agreement.

2. **Applying PLS regression to find the distribution of actual fiber length.**

   For the purpose of improving cotton fiber length measurements and expanding the application of the test results, knowing the actual length distribution is a 'must' and enables a better assessment of the cotton quality. In practice only the projecting fiber length can be measured and the actual fiber length does not. We established an approach to obtaining the actual length distribution from the projecting fiber length. This method is new. Partial least squares regression is the fundamental tool of the

method. The new approach showed promising potential for estimating the original length distribution from the observed length distribution of a fiber beard. Although this method is at beginning stage, we believe it will eventually be adapted by the industry.

3. **A New Distribution.**

To make the PLS regression more efficient a new distribution was introduced, which contains only three parameters, while the mixture of Weibull distributions contains five parameters. It is known that reducing the number of parameters in regression can enhance the estimation of parameters. Therefore, for length by weight, we found that the new distribution fits the data very well. Numerical comparisons for various parameters between the distribution of length from the data and the fitted new distribution indicated that the new distribution works well. PLS was applied on the new distribution to investigate the relationship between the observed length distribution in a fiber beard and the original length distribution.

8

Appendix

Acronyms

AFIS: Advance fiber information system

CV: Coefficient of variance

ML: Mean fiber length by number (AFIS)

ML_w: Mean fiber length by weight (AFIS)

HVI: High volume instrumentation

LHML: Lower-half mean fiber length (HVI)

ML: Mean length

PLS: Partial Least Square

PCA: Principal Component Analysis

SFC: Short fiber content by number (AFIS)

SFC_w: Short fiber content by weight (AFIS)

UHML: Upper-half mean fiber length (HVI)

UI: Uniformity index

UQ: Upper-quartile

USDA: United States Department of Agriculture

Programs

8.1 Mixture of Weibull distribution program

In this section nonlinear regression models were used. Gauss-Newton algorithm and least squares principal to find best fit of fiber length data.

Options $ls = 80$ $ps = 55$ nodate nonumber;

Goption Reset=All;

Symbols to be used in the plots

Symbol1 $v = point$ $l = 1$ $c = black$ $i = spline$;

Symbol2 $v = point$ $l = 2$ $c = red$ $i = spline$;

Symbol3 $v = point$ $l = 1$ $c = green$ $i = spline$;

Symbol4 $v = point$ $l = 2$ $c = bleu$ $i = spline$;

Run;

Import data from excel file

Proc import $datafile = "F : Exeel\ File"$ $out = dat$ $replace$;

Run;

Requests all statistics and tables example Mean Variance etc

Proc univariate data = dat; var length;

Title 'Statistics for length';

Run;

fill is for filling in missing values of length

Data fill; do $i = 0$ *to* 2499;

$Length = i/1000$; $count = 0$; $percent = 0$; $w = .001$;

Output; end; drop i; run;

Data dat1; set dat;

x is rounded to one decimal place. So they are 0, 0.1, 0.2, ...2.5

$Length2 = round(length, .1);$

Run;

Proc freq data=dat1 noprint;

Tables length2/out = out;

Run;

Proc freq data = dat noprint;

Tables length/out = out3;

Run;

Data merg; merge fill out3; by length;

$k = 1$; $y = .01 \times percent/w$;

Run;

Data out; Set out;

$y1 = .01 \times percent/0.1$;

Run;

Proc nlin best = 1 data = out outest = parms;

Parms $p = 0.1$ to 0.5 by 0.1 $a1 = 0.1$ to 5.1 by 1 $b1 = 0.1$ to 5.1 by 1

$a2 = 0.1$ to 5.1 by 1 $b2 = 0.1$ to 5.1 by 1;

Bounds $0 \leq p \leq 1$, $a1 > 0$, $b1 > 0$, $a2 > 0$, $b2 > 0$;

$f1 = a1 * b1 * length2 * *(a1 - 1) * exp(-b1 * length2 * *a1);$

$f2 = a2 * b2 * length2 * *(a2 - 1) * exp(-b2 * length2 * *a2);$

$f = p * f1 + (1 - p) * f2;$

Model $y1 = f$;

Output out = out1 $p = \widehat{y}$;

Title 'compute LS fit to proposed density';

Run;

Proc print data = out;

Title 'output from proc freq with missing values filled in';

Title2 'k is variable for merging, y is proportional to interval prob';

Run;

Legend1 label = none

shape = symbol$(8, 4)$

position = (top right inside)

mode = share;

*Proc gplot data = out*1;

Plot $y1 * length2 = 1\ \widehat{y} * length2 = 2/overlay\ legend = legend1$;

Label $y1 =' data'\ \widehat{y} =' Weibull\ Density'\ length2 =' length'$;

$Title1\ box = 1\ empirical\ and\ fitted\ density$;

$Title2\ h = 2\ color = black\ Projecting\ 30$;

Run;

Data parms; set parms;

If type =' iter';

Run;

Proc sort data = parms; *by descending iter;*

Run;

Compute mean and variance of estimated distribution

Data parms; set parms;

If $N = 1$;

$mu1 = gamma(1 + 1/a1)/b1 * *(1/a1)$;

$sigmasq1 = (gamma(1 + 2/a1) - (gamma(1 + 1/a1)) * *2)/b1 * *(2/a1)$;

$mu2 = gamma(1 + 1/a2)/b2 * *(1/a2)$;

$sigmasq2 = (gamma(1 + 2/a2) - (gamma(1 + 1/a2)) * *2)/b2 * *(2/a2)$;

$mu = p * mu1 + (1-p) * mu2;$

$sigmasq = p * sigmasq1 + (1-p) * sigmasq2;$

$sigma = sqrt(sigmasq);$

$k = 1;$

Keep p a1 b1 a2 b2 mu sigmasq sigma k;

Run;

Merge estimated parameters with original data

Data out;

Merge merg parms; by k;

Keep length count p a1 b1 a2 b2;

Run;

Compute cumulative counts;

Data out; Set out;

If $N = 1$ *then* $cum = count$; *Else* $cum = cum + count$;

Retain cum;

Run;

Compute empirical and proposed distribution functions;

Data out;

Set out;

$pdf2 = (a1 * b1 * length * *(a1-1)) * exp(-b1 * length * *a1);$ $pdf1 = (a2 * b2 * length * *(a2-1)) * exp(-b2 * length * *a2);$

$ecdf = cum/35000;$ $cdf = 1 - p * exp(-b1 * length * *a1) - (1-p) * exp(-b2 * length * *a2);$

$If\ N = 1\ then\ max = abs(ecdf - cdf);$ $else\ max = max(max, abs(ecdf - cdf));$

Retain max; keep length ecdf cdf max pdf1 pdf2;

Run;

Proc gplot data=out;

*Plot pdf*1 ∗ *length* = 1 *pdf*2 ∗ *length* = 2 */overlay legend* = *legend*2;

Run;

Proc gplot data = *out*;

Plot ecdf ∗ *length* = 1 *cdf* ∗ *length* = 2 */overlay legend* = *legend*2;

Title1 box = 1 *'ECDF and Fitted CDF'*;

Title2 h = 2 *'Projecting* 30′;

Run;

Proc gplot data = *new*;

*Plot y*1 ∗ *length*2 = 1 $\widehat{y}$ ∗ *length*2 = 2 *pdf*1 ∗ *length* = 3 *pdf*2 ∗ *length* = 4 *overlay legend* = *legend*1;

*Label y*1 =′ *data*′ $\widehat{y}$ =′ *Fitted Density*′ *length*2 =′ *length*′;

Title1 box = 1 *empirical and fitted density;*

Title2 h = 2 *color* = *black 'Projecting* 30′;

Run;

Quit;

8.2 **Kolmogorov-Smirnov table**

This program generates a Kolmogorov-Smirnov table;

For statistic used see Mood,Graybill & Boes, p.508-509;

The table is easily expanded;

Proc iml;

$m = 0$;

Do x = .4 *to 2.4 by .01;*

$m = m + 1$;

Do i = 1 *to 100;*

If $i = 1$ *then* $sum = 2 * exp(-2 * x * *2)$;

Else $sum = sum + 2 * (-1) * *(i-1) * exp(-2 * i * *2 * x * *2)$;

End;

$p = 1 - sum$;

$u = x \| p$;

If $m = 1$ *then table* $= u$;

Else table $= table/u$;

End;

Create table from table $[colname = x\ hx]$;

Append from table;

Quit;

Proc print data $= table\ split = ' * '$;

Label $hx = '\ asymptotic * probability * Kolmogorov * Smirnov = * sqrt(n) max(d) \leq X'$;

Run;

8.3 **Simulation**

Data dat;

Proc import $datafile = "G : Data\ destination"$ $out = dat$ replace;

Run;

%Let $times = 500$;

%Macro Analysis(da);

*Fill is for filling in missing values of length;

Data fill;

Do $i = 0$ *to 2499; length* $= i/1000$*; count* $= 0$*;*

percent $= 0$; $w = .001$; output;

End; drop i;

Run;

Data dat1;

Set &da;

$Length2 = round(length, .1);$

Run;

Proc freq data = dat1 noprint;

Tables length2 out = out;

Run;

Proc freq data = &da noprint;

Tables length/out = out3;

Run;

Data merg;

Merge fill out3;by length;

$k = 1;$

$y = .01 * percent/w;$

Run;

Data out;

Set out;

$y1 = .01 * percent/0.1;$

Run;

Proc nlin best = 1 data = out outest = parms noprint;

Parms $p = 0.1$ to 0.9 by 0.1

$a1 = 0.1$ to 5.1 by 1

$b1 = 0.1$ to 5.1 by 1

$a2 = 0.1$ to 5.1 by 1

$b2 = 0.1$ to 5.1 by 1;

Bounds $0 \leq p \leq 1, a1 > 0, b1 > 0, a2 > 0, b2 > 0$;

$f1 = a1 * b1 * length2 * *(a1 - 1) * exp(-b1 * length2 * *a1)$;

$f2 = a2 * b2 * length2 * *(a2 - 1) * exp(-b2 * length2 * *a2)$;

$f = p * f1 + (1 - p) * f2$;

Model $y1 = f$;

Output out $= out1$ $p = yhat$;

Title 'compute LS fit to proposed density';

Run;

Data parms;

Set parms;

If $Type =' Iter'$;

Run;

Proc sort data = parms;by descending iter;

Run;

Compute mean and variance of estimated distribution

Data parms;

Set parms;

If $N = 1$;

$mu_1 = gamma(1 + (1/a_1))/b_1 * *(1/a_1)$;

$sigmasq_1 = (gamma(1 + (2/a_1)) - (gamma(1 + (1/a_1))) * *2)/b_1 * *(2/a_1)$;

$mu_2 = gamma(1 + (1/a_2))/b2 * *(1/a_2)$;

$sigmasq_2 = (gamma(1 + (2/a_2)) - (gamma(1 + (1/a_2))) * *2)/b2 * *(2/a_2)$;

$mu = p * mu_1 + (1 - p) * mu_2$;

$sigmasq = p * sigmasq_1 + (1 - p) * sigmasq_2$;

$sigma = sqrt(sigmasq)$;

$k = 1$;

Keep $p\ a1\ b1\ a2\ b2\ mu\ sigmasq\ sigmak$;

Run;

Merge estimated parameters with original data

Data out;

Merge merg parms;by k;

Keep length count p a1 b1 a2 b2;

Run;

Compute cumulative counts

Data out;

Set out;

If $N = 1$ *then* $cum = count$;

Else $cum = cum + count$;

Retain cum;

Run;

Data out;

Set out;

$ecdf = cum/2500$;

$cdf = 1 - p * exp(-b1 * length * *a1) - (1 - p) * exp(-b2 * length * *a2)$;

$d = abs(ecdf - cdf)$;

Run;

Proc means data = out max noprint;

Var d;

Output out = out2 max = ra;

Run;

Data final1;

Set out2;

If $ra * sqrt(2500) < 1.23$ *then* $p1 = 1;$

Else $p1 = 0;$

Run;

Data final2;

Set out2;

$If\ ra * sqrt(2500) < 1.35\ then\ p2 = 1;$

$Else\ p2 = 0;$

Run;

%*Mend Analysis*;

%*Macro permut* (*population*);

%*do* $i = 1$ %*to* & *times*;

Data randomize&i;

Set &*population*; $z = ranuni(0); run;$

Proc sort data = *randomize*&*i*; *byz*;

Run;

Proc surveyselect *data* = &*population method* = *SRS sampsize* = 2500

rep = 1 *seed* = 0 *out* = *sample*&*i*;

Run;

%*End*;

%*Mend permut*;

%*Permut*(*dat*);

%*Macro repetition*;

%*Do* $k = 1$ %*To* ×

%Analysis(*sample*&k);

%*If* &$k = 1$ %*then* %do;

Data FinalCalcul;

Set final1 final2;

Run;

%End;

%Else %do;

Data FinalCalcul;

Set FinalCalcul final1 final2;

Run;

%End;

%End;

%Mend repetition;

%Repetition;

8.4 Partial Least Squares program for variable length

The partial least squares regression method to convert the parameters from the projecting to the original.

Ods pdf $(id = fancy)$ $file =$ "$J :$ *Newdata output8cot.pdf*";

Options $ls = 80$ $ps = 55$ nodate nonumber;

Data Fiber;

Input n aOrig b1Orig c1Orig b2Orig c2Orig aProj b1Proj c1Proj b2Proj c2Proj;

Datalines;

'Include the data'

Data Fiber; *set path*; *Run;*

Proc pls data = patha $nfac = 6$;

Model aOrig b1Orig c1Orig b2Orig c2Orig $=$ *aProj b1Proj c1Proj b2Proj c2Proj/intercept*;

$Output\ out = outpls\ predicted = yhat1 - yhat5$

$yresidual = yres1 - yres5$

$xresidual = xres1 - xres5$

$xscore = xscr$

$yscore = yscr;$

$Run;$

$\%Let\ ifac = 1;$

$Data\ cotton;\ set\ outpls;$

$Length\ text\ 2;$

$Retain\ function\ 'label'\ position\ '5'\ hsys\ '3'\ xsys\ '2'\ ysys\ '2'$

$color\ 'blue'\ style\ 'swissb';$

$Text = \%str(n);\ x = xscr\&ifac;\ y = yscr\&ifac;$

$Axis1\ label = (angle = 270 rotate = 90"Yscore\&ifac")$

$major = (number = 5) minor = none;$

$Axis2\ label = ("X - score\&ifac") minor = none;$

$symbol1v = none i = none;$

$Proc\ gplot\ data = outpls;$

$Plot\ yscr\&ifac * xscr\&ifac = 1$

$Anno{=}cotton\ vaxis = axis1\ haxis = axis2$ frame $cframe = ligr;$

$Run;$

$Ods\ output\ XWeights = xweights;$

$Proc\ pls\ data = patha\ nfac = 6$ details;

$Model\ aOrig\ b1Orig\ c1Orig\ b2Orig\ c2Orig\ =\ aProj\ b1Proj\ c1Proj\ b2Proj\ c2Proj / intercept;$

$Run;$

$Proc\ transpose\ data = xweights(drop = Number\ Of\ Factors\ Inner\ Reg\ Coef)$

$out = xweights;$

Data xweights; set xweights;

*Rename col*1 = *w*1 *col*2 = *w*2 *col*3 = *w*3;

Data wt; set xweights;

Length text $ 2;

Retain function '*label*' *position* '5' hsys '3' xsys '2' ysys '2' *color* '*blue*' *style* '*swissb*';

$Text = \%str(\_name\_); x = w1; y = w2;$

Run;

Axis1 label = (*angle* = 270*rotate* = 90"*Xweight*2")

major = (*number* = 5)*minor* = *none*;

Axis2 label = ("*X weight* 1")*minor* = *none*;

Symbol1 v = *none i* = *none*;

Proc gplot data = *xweights*;

*Plot w*2 ∗ *w*1 = 1 *anno* = *wtanno vaxis* = *axis*1

haxis = *axis*2 frame *cframe* = *ligr*;

Run; quit;

Ods listing close;

Ods output PercentVariation = *pctvar*

XWeights = *xweights*

CenScaleParms = *solution*;

Proc pls data = *patha nfac* = 6 *details*;

*Model aOrig b*1*Orig c*1*Orig b*2*Orig c*2*Orig* = *aProj b*1*Proj c*1*Proj b*2*Proj c*2*Proj intercept*;

Run;

Ods listing;

Transpose weights and R**2' s

Data xweights; *set xweights*; *name* =′ *W*′|||*trim*(*left*(*_n_*));

$Data\ \ pctvar\ \ ;\ \ set\ \ pctvar\ \ ;\ \ name =' R'||trim(left(\_n\_));$

$Proc\ \ transpose\ \ data = xweights\ (drop = Number\ Of\ \ Factors\ \ InnerRegCoef)$

$Out = xweights;$

$Proc\ transpose\ data = pctvar(keep = \_name\_CurrentYVariation)$

$Out = pctvar;$

Run;

Proc sql;

Create table vip as

Select $*,$

$w1sqrt(uss(w1))\ as\ wnorm1,$

$w2sqrt(uss(w2))$ a$s\ wnorm2$

$From\ \ xweights\ \ left\ \ join\ \ pctvar(drop = \_name\_)on1;$

Data vip; set vip; keep name vip;

Array $wnorm2;$

Array r2;

$VIP = 0;$

$Do\ i = 1to2;$

$VIP = VIP + ri * (wnormi * *2)/sum(ofr1 - r2);$

End;

$VIP = sqrt(VIP * 4);$

$Data$ vipbpls; merge solution $vip(drop = name);$

$Proc\ print\ data = vipbpls;$

Run;

$Proc$ pls $data = patha\ \ cv = random;$

$Model\ \ aorig\ b1orig\ c1orig\ b2orig\ c2orig\ =\ aProj\ b1Proj\ c1Proj\ b2Proj\ c2Proj\ intercept;$

Run;

Proc pls *data* = *patha cv* = *random cvtest*(*seed* = 12345);

*Model aorig b*1*orig c*1*orig b*2*orig c*2*orig* = *aProj b*1*Proj c*1*Proj b*2*Proj c*2*Projintercept*;

Run;

Data newobs;

Input n $ aProj b1Proj c1Proj b2Proj c2Proj ;

Datalines;

Data all; set patha newobs;

Proc pls data = *all nfac* = 5;

*Model aOrig b*1*Orig c*1*Orig b*2*Orig c*2*Orig* = *aProj b*1*Proj c*1*Proj b*2*Proj c*2*Proj intercept*;

Output out = *pred p* = *paorig pb*1*orig pc*1*orig pb*2*orig pc*2*orig*;

Proc print data = *pred*;

Where (*nin*('8'));

Var n paorig pb1Orig pc1Orig pb2Orig pc2Orig;

Run;

Ods pdf (*id* = *fancy*) close;

Quit;

9

References

1. Agnar H.Skuldsson 1988. PLS Regression. Methods. *Journal of Chemometrics* 2,211-228

2. Bent Jrgensen Yuri Goegebeur. Master Of Applied Statistics, Department of Statistics, *University of Southern Denmark*, 2002-06. available at./http://statmaster.sdu.dk/courses

3. Cai, Y., Cui, X., Belmasrour, R., Li, L., Delhom, C.D., Rodgers III, J.E., Martin, V., Watson, M. 2010. Using Partial Least Squares Regression to Obtain Cotton Fiber Length Distributions From The Beard Testing Method. Proceeding of the 2010 National Cotton Council Beltwide Cotton Conference, January 5-7, 2010, New Orleans, Louisiana. p. 1411-1413. 2010 CDROM.

4. Cai, Y., Cui, X., Rodgers, J., Martin, V., and Watson, M. 2009. An Investigation on the Sampling Bias of the Beard Method as Used in HVI, *Journal of the Textile Institute*,in press.

5. Chu, Y.T. and Riley C.R. 1997. New interpretation of the fibrogram, *Textile Research Journal*. 67-897-901.

6. Cui, X., Calamari, T.A. and Robert, K.Q. Jr.1997. An investigation of cotton fiber lengths measured by HVI and AFIS. *The 10th EFS system research forum proceedings*.115-123.

7. Cui, X., T.A. Calamari, and M. Suh. 1998. Theoretical and practical aspects of fiber length comparisons of various cottons. *Textile Research Journal*. 68-467-472.

8. Cui, X., D. Thibodeaux, K.Q. Robert, Jr., and T.A. Calamari. 2004. Statistical Parameters of Cotton Short Fibers. Proceedings of the *Beltwide Cotton Conference*. 2383-2386.

9. Cui, X. Rodgers, J. Cai, Y. Li, L. and Belmasrour, R. 2009. Obtaining Cotton Fiber Length Distributions from the Beard Test Method, Part 1 - Theoretical Distributions Related to the Beard Method. *The Journal of Cotton*, 13-265-273.

10. Electronic Textbook StatSoft available at. /www.statsoft.com/textbook/stpls.html

11. Garthwaite, Paul H. 1994. An Interpretation of Partial Least Squares. *Journal of the American Statistical Association*. 89- 123-127.

12. High volume tests of staple fiber length. *Spinlab Technical Report*. January 1981.

13. Hertel, K.L. 1936. An optical method for the length analysis of cotton fibers. *Textile Research Journal*. 6:331-339 .

14. Hertel, K.L. 1940. A method of fiber-length analysis using the fibrograph. *Textile Research Journal*. 10-510-525.

15. Krifa, M. 2006. Fiber Length Distribution in Cotton Processing: Dominant Features and Interaction Effects. *Textile Research Journal*. 76-426-435.

16. Krifa, M. 2007 Multiparameter Comparison of Cotton Fiber Length Distribution. World Cotton Research Conference - 4, Lubbock, TX, USA.

17. Krifa, M. 2008. Fiber Length Distribution in Cotton Processing: A Finite Mixture Distribution Model. *Textile Research Journal*. 78-688-698.

18. Krowicki, R.S., D.P. Thibodaux, and Duckett, K.E. 1996. Generating fiber length distribution from the fibrogram. *Textile Research Journal*. 66-306-310.

19. Mood, A.M., F.A. Graybill, and Boes, D.C. 1974 Introduction to the theory of statistics, *3rd edition*, McGraw-Hill.

20. Prier, H.W. and Sasser, P.E. 1972. The mathematical basis of fiber-length analysis from fibrogram data. *Textile Research Journal*. 42-410-419.

21. Robert, K.Q. 2005, Cotton fiber breakage and its relation to length distribution, short fiber, and uniformity, Proceedings of *the Beltwide Conference*, 3074-86.

22. Randall D. Tobias. An Introduction to Partial Least Squares Regression. , SAS Institute Inc., Cary, NC

23. Schorack, G.R. and Wellner, J.A. 1986. Empirical Processes with Applications to Statistics, New York: John Wiley and Sons.

24. Suh, M., and Sasser, P.E. 1996 Technological and Economic Impact of HVI on Cotton and Cotton Textile Industries. *Journal of the Textile Institute* Part 3, 87-43-59.

25. Woo, J.L. 1967. An Appraisal of the Length Measures Used for Cotton Fibres. *Journal of the Textiles Institute*, 58(11), 557-572.

26. Wold S, 1994 PLS for Multivariate Linear Modeling QSAR: Chemometric Methods in Molecular Design. Methods and Principles in Medicinal Chemistry van de Water-beemd H (Editor) Verlag- Chemie.

27. Wold, Svante. PLS for Multivariate Linear Modeling, March, 1986. Partial Least Squares (PLS)

28. Zeidman, M.I., Batra, S.K. and Sasser, P.E. 1991. Determining short fiber content in cotton. *Textile Research Journal*. 61-21-30.

10

Vita

Rachid Belmasrour was born on March 17, 1973, in Casablanca, Morocco. He obtained his B.S. degree in Applied Mathematics in 1998 from University of Hassan II, Casablanca, and his M.S degree in Mathematical Informatics in 2001 from University of Versailles Saint Quentin. In 2004 He moved to the United States of America.

In the fall of 2005, he was admitted to the graduate school of the University of New Orleans to pursue studies in statistics toward the PhD. degree. From January 2007, he was employed by the United State Department of Agriculture (USDA), Division of Cotton Structure and Quality (CSQ) as a research student assistant.