FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS AND STATISTICS

END OF SEMESTER EXAMINATION PAPER 2015

MODULE CODE: MA4605

SEMESTER: Autumn 2015

MODULE TITLE: Chemometrics

DURATION OF EXAM: 2.5 hours

LECTURER: Mr. Kevin O'Brien

GRADING SCHEME: 100 marks
60% of module grade

EXTERNAL EXAMINER: Prof. J. King

INSTRUCTIONS TO CANDIDATES

Scientific calculators approved by the University of Limerick can be used.
Formula sheet and statistical tables areprovided at the end of the exam paper.
Students must attempt any 4 questions from 5.

Question 1. (25 marks) Inference Procedures and Distributional Testing

Question 1 Part A (15 Marks)

A test of a specific blood factor has been devised such that, for adults in Western Europe, the test score is normally distributed with mean 100 and standard deviation 10. A clinical research organization is carrying out research on the blood factor levels for sufferers of a particular disease.

- A study has obtained the following test scores for 14 randomly selected patients suffering from the disease in Ireland

$$\{118, 116, 109, 105, 103, 111, 139, 117, 107, 105, 125, 99, 106, 103\}$$

- A similar study has obtained the following test scores for 14 randomly selected patients suffering from the disease in Denmark.

$$\{120, 140, 112, 109, 114, 116, 99, 108, 109, 111, 109, 131, 117, 101\}$$

The following blocks of `R` code (i.e. blocks A to E) are based on the data for this assessment. Write a short report on your conclusion for this assessment.

Marking Scheme*: Either 2 or 3 Marks will be awarded for a correct interpretation of each code segment, for a total of 12 Marks. Remember to state the null and alternative hypotheses when relevant. 3 Marks will also be award for an overall conclusion.*

Block A (3 Marks)

```
> grubbs.test(X)

Grubbs test for one outlier

data:  X
G = 2.56990, U = 0.45291, p-value = 0.01748
alternative hypothesis: highest value 139 is an outlier

> grubbs.test(Y)

Grubbs test for one outlier

data:  Y
G = 2.17410, U = 0.66387, p-value = 0.1486
alternative hypothesis: highest value 131 is an outlier
```

Block B (2 Marks)

```
> shapiro.test(X)

Shapiro-Wilk normality test

data:  X
W = 0.87633, p-value = 0.05153



> shapiro.test(Y)

Shapiro-Wilk normality test

data:  Y
W = 0.92341, p-value = 0.1914
```

Block C (2 Marks)

```
> var.test(X,Y)

F test to compare two variances

data:  X and Y
F = 2.3808, num df = 13, denom df = 15, p-value = 0.1107
alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:
0.8139616 7.2677776

sample estimates:
ratio of variances
2.38076
```

Block D (3 Marks)

```
> t.test(X,Y,var.equal=TRUE)

Two Sample t-test

data:  X and Y
t = -1.3471, df = 28, p-value = 0.1888
alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
-10.982656   2.268371

sample estimates:
mean of x mean of y
111.6429  116.0000
```

```
> t.test(X,Y,var.equal=FALSE)

Welch Two Sample t-test

data:  X and Y
t = -1.3096, df = 21.764, p-value = 0.204
alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
-11.261465   2.547179

sample estimates:
mean of x mean of y
111.6429  116.0000
```

Block E

```
> wilcox.test(X,Y)

Wilcoxon rank sum test with continuity correction

data:  X and Y
W = 68, p-value = 0.06999
alternative hypothesis: true location shift is not equal to 0
```

Question 1 Part B (4 Marks)

Numeric Transformations, such as logarithmic transformation, are often used in statistical analysis as an approach for dealing with non-normal data.

(i.) (1 Marks) Describe the purpose of Tukey's Ladder (referencing direction and relative strength).

(ii.) (2 Marks) Give two examples of a transformation for various types of skewed data (i.e. an example for both types of skewness).

(iii.) (1 Marks) Discuss the limitations of numeric transformations.

Question 1 Part C (6 Marks)

(i.) (3 Marks) Provide a brief description for three tests from the family of Grubb's Outliers Tests. Include in your description a statement of the null and alternative hypothesis for each test, any required assumptions and the limitations of these tests.

(ii.) (3 Marks) Showing your working, use the Dixon Q Test to test the hypothesis that the maximum value of the following data set is an outlier.

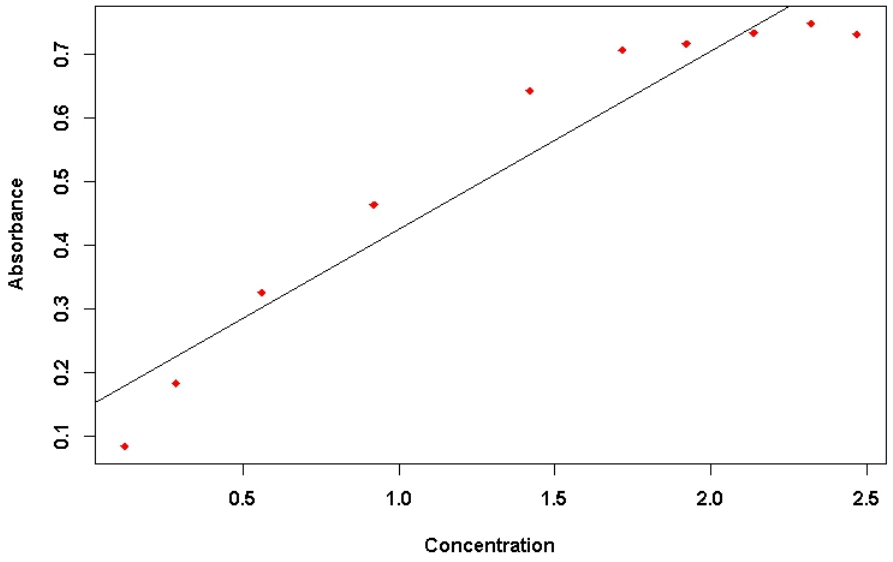
$$19, 22, 23, 24, 25, 26, 29, 38$$

Question 2. (25 marks) Regression Models

Question 2 Part A (8 Marks)

In an experiment to determine hydrolysable tannins in plants by absorption spectroscopy, the following results from ten samples were obtained and are tabulated below. A simple linear regression model, predicting absorbance values using concentration as the independent variable, was fitted to the data. The scatterplot is depicted below.

Sample	1	2	3	4	5
Absorbance	0.084	0.183	0.326	0.464	0.643
Concentration	0.123	0.288	0.562	0.921	1.420
Sample	6	7	8	9	10
Absorbance	0.707	0.717	0.734	0.749	0.732
Concentration	1.717	1.921	2.137	2.321	2.467



(i.) (1 marks) Is the simple linear regression model approach suitable for this study? Explain your answer with reference to the scatter-plot.

(ii.) (3 marks) Two polynomial models were also fitted to the data. Description of all three fitted models are found in the three blocks of `R` code on the following pages. The *Akaike information criterion* is listed, for each of the three fitted models. Write down the regression equations of each of the three models.

(iii.) (2 marks) Specify which one of the models you would use. Justify your answer with appropriate statistical values.

(iv.) (2 marks) Using the best fitting model, predict a value for absorbance when the concentration level is 1.2 mg/ml.

Model 1

```
> summary(Model1)
Call:
lm(formula = Absorb ~ Conc)
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.14412    0.04721   3.053   0.0158 *
Concentration  0.28088    0.02930   9.586 1.16e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.07584 on 8 degrees of freedom
Multiple R-squared: 0.9199,     Adjusted R-squared: 0.9099
F-statistic: 91.89 on 1 and 8 DF,  p-value: 1.163e-05
>
>
>AIC(Model1)
[1] -19.4343
```

Model 2

```
> summary(Model2)
Call:
lm(formula = Absorb ~ Conc + Conc.Squared)
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.006582   0.008013   0.821    0.439
Concentration  0.642935   0.015568  41.299 1.27e-09 ***
Conc.Squared  -0.140573   0.005894 -23.851 5.79e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.008939 on 7 degrees of freedom
Multiple R-squared: 0.999,      Adjusted R-squared: 0.9987
F-statistic:  3592 on 2 and 7 DF,  p-value: 2.879e-11
>
>
> AIC(Model2)
[1] -61.5338
```

Model 3

```
> summary(Model3)

Call:
lm(formula = Absorb ~ Conc+ Conc.Squared + Conc.Cubed)
...
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.013712   0.011629   1.179   0.2830
Concentration  0.608682   0.042825  14.213 7.58e-06 ***
Conc.Squared  -0.108186   0.038088  -2.840   0.0296 *
Conc.Cubed    -0.008196   0.009518  -0.861   0.4223
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.009109 on 6 degrees of freedom
Multiple R-squared: 0.9991,     Adjusted R-squared: 0.9987
F-statistic:  2306 on 3 and 6 DF,  p-value: 1.422e-09
>
>
> AIC(Model3)
[1] -60.69903
```

Question 2 Part B (7 Marks)

In certain circumstances, Robust Regression may be used in preference to Ordinary Least Squares Regression. Answer the following questions relating to Robust Regression.

(i.) (1 Mark) Describe what these circumstances might be.

(ii.) (1 Mark) State one difference between OLS and Robust regression techniques, in terms of computing regression equations.

(iii.) (2 Marks) Explain the process of Huber Weighting for Residuals, stating the algorithm used to compute weightings.

(iv.) (3 Marks) Suppose that Huber Weighting, with a tuning constanct of $k = 13.45$, was applied to the observations tabulated below. What would be the outcome of the procedure for each case.

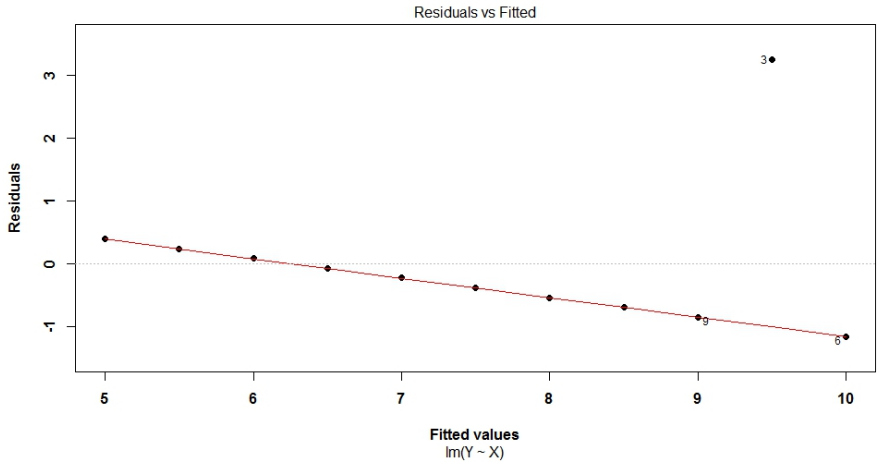
Observation i	Residual e_i
11	-9.07
14	14.54
18	22.91

Question 2 Part C (10 Marks)

(i.) (1 Marks) Explain the term "Influence" in the context of linear regression models. Support your answer with sketches.

(ii.) (1 Marks) Explain the term "Cook's Distance" in the context of linear regression models.

(iii.) (2 Marks) The following plot is the *Residual vs Fitted* plot, the first of `R`'s diagnostic plots for linear models. Briefly describe how to interpet this plot. What is your conclusion?



(iv.) (2 Marks) Explain the term "Heteroscedascity" in the context of linear regression models. Support your answer with sketches.

(v.) (1 Mark) The Non-constant Variance Score Test was carried out to test for Heteroscedascity. The output is depicted below. State your conclusion to the following procedure.

```
> ncvTest(ModelQ2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 7.585432    Df = 1     p = 0.005884187
```

(vi.) (2 Marks) The Durbin Watson Test was carried out to test for Autocorrelation. Briefly describe autocorrelation. You may support your answer with sketches.

(vii.) (1 Mark) State your conclusion to the following procedure.

```
> durbinWatsonTest(ModelQ2)
lag Autocorrelation D-W Statistic p-value
1     -0.08428163      2.143578   0.806
Alternative hypothesis: rho != 0
```

Question 3. (25 marks) Experimental Design

Question 3 Part A (15 Marks)

Three investigators, A, B and C, performed six determination of nitrate in water using the same procedure. The results in μM were:

A	B	C
6.7	6.3	6.8
6.8	6.2	6.9
6.5	6.1	7.1
6.8	6.3	6.9
6.9	6.5	7.2
7.1	6.4	7.1

We are also given the summmary statistics for each of the three investigators, as well as for the samples combined.

	Sample Mean	Sample Variance
A	6.8	0.040
B	6.3	0.020
C	7	0.024
Overall	6.7	0.1164

An analysis of variance procedure is used to determine if there is a significant difference between the mean of the determinations made by the three investigators.

The following questions will result in the completion of the ANOVA Table on the next page. The $p-$value is already provided.

(i.) (3 Marks) Compute the Between Groups Sum of Squares. (Show your workings.)

(ii.) (3 Marks) Compute the Within Groups Sum of Squares.(Show your workings.)

(iii.) (2 Marks) Compute the Total Sum of Squares. (Show your workings).

(iv.) (2 Marks) State the degrees of freedom for the ANOVA Tables

(v.) (1 Marks) Compute the Mean Square values.

(vi.) (1 Marks) Compute the test Statistic for this procedure (i.e. the F-value.)

(vii.) (3 Marks) This analysis is used to assess if there is any difference between the mean determinations made by the three investigators. What is your conclusion? Clearly state the null and alternative hypothesis.

Source	DF	SS	MS	F	p-value
Between	?	?	?	?	8.9×10^{-06}
Within	?	?	?		
Total	?	?			

Question 3 Part B (4 Marks)

The R code and graphical procedures, below and on the following page, are relevant to checking whether the underlying assumptions are met for an ANOVA model.

(i.) (2 marks) State two testable assumptions required for ANOVA procedures? (You may refer to the code output below.)

(ii.) (2 marks) Assess the validity of these assumptions for an ANOVA model based on the following code outputs.

```
Shapiro-Wilk normality test

data:  Residuals
W = 0.9719, p-value = 0.3819
```

```
Bartlett test of homogeneity of variances

data:  Experiment
Bartlett's K-squared = 105.9585, df = 1, p-value < 2.2e-16
```

Question 3 Part C (6 Marks)

Suppose you want to determine whether the brand of cleaning product used and the temperature affects the amount of dirt removed from your machinery. You are also interested in determining if there is an interaction between the two variables.

You buy two different brand of detergent ("*Super*" and "*Best*") and choose three different temperature levels ("*Cold*", "*Warm*", and "*Hot*"). There are four measurements per treatment group.

	Cold	Warm	Hot
Super	4,5,6,5	7,9,8,12	10,12,11,9
Best	6,6,4,4	13,15,12,12	12,13,10,13

- Detergent is Factor A.
- Temperature is Factor B.
- The variance of the response variable is 12.2011.

Source	DF	SS	MS	F
A	?	22.04	?	?
B	?	?	102.37	?
A:B	?	16.08	?	?
Resid	?	?	?	
Total	?	?		

For the table above, replace the questions marks with the correct values in each of the following columns. (The number of marks for each column is indicated here:)

(i.) (2 Marks) Degrees of freedom

(ii.) (2 Mark) Sums of Squares column

(iii.) (1 Mark) Mean Square Values

(iv.) (1 Mark) F-Values

Question 4. (25 marks) Experimental Design

Question 4 Part A (25 Marks)

In an investigation into the extraction of nitrate-nitrogen from air dried soil, three quantitative variables were investigated at two levels. These were the amount of oxidised activated charcoal (A) added to the extracting solution to remove organic interferences, the strength of CaSO4 extracting solution (C), and the time the soil was shaken with the solution (T). The aim of the investigation was to optimise the extraction procedure. The levels of the variables are given here:

		-	+
Activated charcoal (g)	A	0.5	1
CaSO4 (%)	C	0.1	0.2
Time (minutes)	T	30	60

The results are given below and are the amounts recovered (expressed as the percentage of known nitrate concentration).

A	C	T	y		
-1	-1	-1	45.1	44.6	45.7
1	-1	-1	44.9	45.3	44.1
-1	1	-1	45.8	46.7	46.3
1	1	-1	43.7	43.8	44.3
-1	-1	1	33.3	32.3	34.1
1	-1	1	51.7	53.8	52.1
-1	1	1	32.6	31.8	34.1
1	1	1	52.2	53.2	51.3

(i.) (7 Marks) Calculate the contrasts.

(ii.) (3 Marks) Calculate the effects.

(iii.) (3 Marks) Calculate the sum of squares for the ANOVA Table.

(iv.) (4 Marks) Using the computed sums of squares values, complete the ANOVA table (see the `R` code below).

(v.) (4 Marks) Comment on the tests for significant for the main effects and interactions. State clearly your conclusions.

(vi.) (4 Marks) Write down a Regression equation that can be used predicting amounts based on the results of this experiment.

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
A	...	...	...	...	$7.39e^{-15}$ ***
B	...	...	...	...	0.960
C	...	...	...	...	$3.92e^{-06}$ ***
A:B	...	...	...	...	0.257
A:C	...	...	...	...	$6.25e^{-16}$ ***
B:C	...	...	...	...	0.322
A:B:C	...	...	...	...	0.203
Residuals	...	...			
Total	...	1172.985			

Question 5. (25 marks) Statistical Process Control

Question 5 Part A (6 Marks)

A normally distributed quality characteristic is monitored through the use of control charts. These charts have the following parameters. All charts are in control.

	LCL	Centre Line	UCL
$\bar{X}$-Chart	995	1000	1005
R-Chart	0	21	44.394

(i.) (2 Marks) What sample size is being used for this analysis?

(ii.) (2 Marks) Estimate the mean of the process standard deviations $\bar{s}$.

(iii.) (2 Marks) Compute the control limits for the process standard deviation chart (i.e. the s-chart).

Question 5 Part B (7 Marks)

An automobile assembly plant concerned about quality improvement measured sets of five camshafts on twenty occasions throughout the day. The specifications for the process state that the design specification limits at 600±3mm.

(i.) (4 Marks) Determine the *Process Capability Indices* C_p and C_{pk}, commenting on the respective values. Use the `R` code output on the following page.

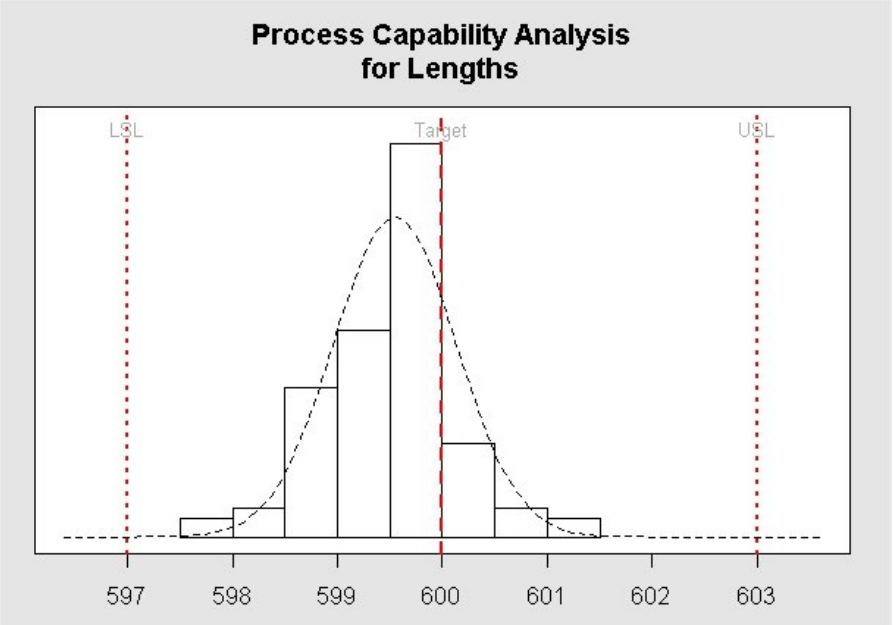
(ii.) (2 Mark) Explain why there would be a discrepancy between C_p and C_{pk}. Illustrate your answer with sketches.

(iii.) (1 Mark) Comment on the graphical output of the *Process Capability Analysis*, also presented on the next page.

```
Process Capability Analysis

Call:
process.capability(object = obj,
          spec.limits = c(597, 603))
Number of obs = 100          Target = 600
Center = 599.548         LSL = 597
StdDev = 0.5846948       USL = 603

Capability indices:
      Value   2.5%  97.5%
Cp    ...
Cp_l  ...
Cp_u  ...
Cp_k  ...
Cpm   1.353  1.134  1.572
Exp<LSL 0%   Obs<LSL 0%
```



Question 5 Part C (12 Marks)

The **Nelson Rules** are a set of eight decision rules for detecting "out-of-control" or non-random conditions on control charts. These rules are applied to a control chart on which the magnitude of some variable is plotted against time. The rules are based on the mean value and the standard deviation of the samples.

(i) (4×3 Marks) Discuss any four of these rules, and how they would be used to detect "out of control" processes. Support your answer with sketch.

In your answer, you may make reference to the following properties of the Normal Distribution. Consider the random variable X distributed as

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

where μ is the mean and σ^2 is the variance of an random variable X.

- $\Pr(\mu - 1\sigma \leq X \leq \mu + 1\sigma) = 0.6827$
- $\Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9545$
- $\Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.9973$

Critical Values for Dixon Q Test

N	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
3	0.941	0.970	0.994
4	0.765	0.829	0.926
5	0.642	0.710	0.821
6	0.560	0.625	0.740
7	0.507	0.568	0.680
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568
11	0.392	0.444	0.542
12	0.376	0.426	0.522
13	0.361	0.410	0.503
14	0.349	0.396	0.488
15	0.338	0.384	0.475
16	0.329	0.374	0.463

Two Way ANOVA

$$MS_A = c \times S_r^2$$
$$MS_B = r \times S_c^2$$

Control Limits for Control Charts

$$\bar{\bar{x}} \pm 3\frac{\bar{s}}{c_4\sqrt{n}}$$

$$\bar{s} \pm 3\frac{c_5\bar{s}}{c_4}$$

$$\left[\bar{R}D_3, \bar{R}D_4\right]$$

Process Capability Indices

$$\hat{C}_p = \frac{\text{USL} - \text{LSL}}{6s}$$

$$\hat{C}_{pk} = \min\left[\frac{\text{USL} - \bar{x}}{3s}, \frac{\bar{x} - \text{LSL}}{3s}\right]$$

$$\hat{C}_{pm} = \frac{\text{USL} - \text{LSL}}{6\sqrt{s^2 + (\bar{x} - T)^2}}$$

2^3 Design: Interaction Effects

$$AB = \frac{1}{4n}\left[abc - bc + ab - b - ac + c - a + (1)\right]$$

$$AC = \frac{1}{4n}\left[(1) - a + b - ab - c + ac - bc + abc\right]$$

$$BC = \frac{1}{4n}\left[(1) + a - b - ab - c - ac + bc + abc\right]$$

$$ABC = \frac{1}{4n}\left[abc - bc - ac + c - ab + b + a - (1)\right]$$

Factorial Design: Sums of Squares

$$\text{Effect} = \frac{\text{Contrast}}{4n}$$

$$\text{Sums of Squares} = \frac{(\text{Contrast})^2}{8n}$$

Factors for Control Charts

Sample Size (n)	c4	c5	d2	d3	D3	D4
2	0.7979	0.6028	1.128	0.853	0	3.267
3	0.8862	0.4633	1.693	0.888	0	2.574
4	0.9213	0.3889	2.059	0.88	0	2.282
5	0.9400	0.3412	2.326	0.864	0	2.114
6	0.9515	0.3076	2.534	0.848	0	2.004
7	0.9594	0.282	2.704	0.833	0.076	1.924
8	0.9650	0.2622	2.847	0.82	0.136	1.864
9	0.9693	0.2459	2.970	0.808	0.184	1.816
10	0.9727	0.2321	3.078	0.797	0.223	1.777
11	0.9754	0.2204	3.173	0.787	0.256	1.744
12	0.9776	0.2105	3.258	0.778	0.283	1.717
13	0.9794	0.2019	3.336	0.770	0.307	1.693
14	0.9810	0.1940	3.407	0.763	0.328	1.672
15	0.9823	0.1873	3.472	0.756	0.347	1.653
16	0.9835	0.1809	3.532	0.750	0.363	1.637
17	0.9845	0.1754	3.588	0.744	0.378	1.622
18	0.9854	0.1703	3.64	0.739	0.391	1.608
19	0.9862	0.1656	3.689	0.734	0.403	1.597
20	0.9869	0.1613	3.735	0.729	0.415	1.585
21	0.9876	0.1570	3.778	0.724	0.425	1.575
22	0.9882	0.1532	3.819	0.720	0.434	1.566
23	0.9887	0.1499	3.858	0.716	0.443	1.557
24	0.9892	0.1466	3.895	0.712	0.451	1.548
25	0.9896	0.1438	3.931	0.708	0.459	1.541