FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS AND STATISTICS

END OF SEMESTER EXAMINATION PAPER 2016

MODULE CODE: MA4605

SEMESTER: Autumn 2016

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. A. Marshall

INSTRUCTIONS TO CANDIDATES

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

Question 1. (25 marks) Inference Procedures

(a) The concentration of nicotine was determined by gas chromatography in each of two samples of known concentrations 10 ng/ml and 50 ng/ml.

```
Data: Sample (Lo): m = 10 ng/ml, n=14.

8.40, 9.59, 9.38, 9.10, 10.78, 11.41, 9.94,
10.08, 12.11, 9.10, 9.59, 10.36, 10.41, 10.52.

Data: Sample (Hi): m = 50 ng/ml, n=10.

47.5, 48.4, 48.8, 48.4, 46.8,
46.2, 48.6, 50.6, 45.5, 46.1.
```

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.

(i) (12 Marks) State the purpose of each code segment and interpret it. Remember to state the null and alternative hypotheses for each segement.

(ii) (3 Marks) Provide an overall conlcusion to the analysis.

Block A (2 Marks)

```
> grubbs.test(Lo)

Grubbs test for one outlier

data:  Lo
G = 2.09700, U = 0.63573, p-value = 0.1561
alternative hypothesis: highest value 12.11 is an outlier


> grubbs.test(Hi)

Grubbs test for one outlier

data:  Hi
G = 1.86510, U = 0.57054, p-value = 0.1985
alternative hypothesis: highest value 50.6 is an outlier
```

Block B (2 Marks)

```
> var.test(Hi,Lo)
F test to compare two variances

data:  Lo and Hi
F = 0.3945, num df = 13, denom df = 9, p-value = 0.1246

alternative hypothesis:
true ratio of variances is not equal to 1

95 percent confidence interval:
0.1029905 1.3066461

sample estimates:
ratio of variances
0.3945149
```

Block C (2 Marks)

```
> shapiro.test(Lo)

Shapiro-Wilk normality test

data:  Lo
W = 0.9779, p-value = 0.9609


> shapiro.test(Hi)

Shapiro-Wilk normality test

data:  Hi
W = 0.9496, p-value = 0.6634
```

Block D (3 Marks)

```
> t.test(Lo,Hi)

Welch Two Sample t-test

data:  Lo and Hi
t = -67.374, df = 14.016, p-value < 2.2e-16

alternative hypothesis:
true difference in means is not equal to 0

95 percent confidence interval:
-38.83294 -36.43706
sample estimates:
mean of x mean of y
10.055    47.690
```

Block E (3 Marks)

```
> t.test(Lo,Hi,var.equal=TRUE)

Two Sample t-test

data:  Lo and Hi
t = -72.6977, df = 22, p-value < 2.2e-16

alternative hypothesis:
true difference in means is not equal to 0

95 percent confidence interval:
-38.70863 -36.56137
sample estimates:
mean of x mean of y
10.055    47.690
```

(b) (5 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 individuals with a particular disease, with emphasis on the effects of medication on the blood factor level.

For a group of 10 volunteer patients the following test scores were obtained both prior to, and after the medication.

Patient	A	B	C	D	E	F	G	H	I	J
Before	120	140	112	109	114	116	99	108	109	111
After	104	112	110	107	101	103	101	102	103	102

The organization wishes to determine if there is a significant improvement (lessening of the blood factor level) due to the medication. Using the output, shown below, write a short report discussing your findings. State the null and alternative hypotheses clearly. (You may assume that the case-wise differences are normally distributed.)

```
> t.test(Before,After,paired=TRUE)

Paired t-test

data:  Before and After
t = 3.3881, df = 9, p-value = 0.008023
alternative hypothesis:
  true difference in means is not equal to 0

95 percent confidence interval:
3.090618 15.509382
sample estimates:
mean of the differences
9.3
```

(c) (5 Marks) The organization wishes to assess the link between the blood factor level and intake of a particular vitamin. Using the following output, write a short report discussing your findings. State the null and alternative hypotheses clearly.

(The R output is presented on the next page)

```
> cor.test(BloodFactor,Vitamin)

Pearson's product-moment correlation

data:  BloodFactor and Vitamin
t = -3.3672, df = 8, p-value = 0.009827

alternative hypothesis:
true correlation is not equal to 0

95 percent confidence interval:
-0.9414504 -0.2627543
sample estimates:
cor
-0.7657047
```

Question 2. (25 marks) Regression Models

(a) Two methods of determing sulphide content exist: ios-selective estrode (ISE) and gravimetry. Experimental results obtained by both methods are expressed in milligrams of sulphide, and are tabulated below.

Sample Number	1	2	3	4	5	6	7	8	9	10
ISE method	108	12	152	3	106	11	128	12	160	128
gravimetry	105	16	113	1	108	11	141	161	182	118

Two simple linear models are fitted to the data. Model 1 uses the gravimetric measurement as an independent variable to predict the ISE measurement. Conversely, Model 2 uses the ISE measurement as an independent variable used to predict the gravimetric measurement.

(The **R** *output is presented on the next page)*

Model 1

```
Call:
lm(formula = ISE ~ grav)
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.1125    28.8487   0.524    0.615
grav          0.6997     0.2543   2.751    0.025 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
....
```

Model 2

```
Call:
lm(formula = grav ~ ISE)
..
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.6215    25.8542   1.494    0.174
ISE           0.6949     0.2526   2.751    0.025 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
....
```

(i) (3 Marks) Write down the regression equation for both of the fitted models. Briefly comment on the significant of each of the regression estimates.

(ii) (3 Marks) Is a simple linear regression model a suitable approach for this type of analysis? Explain why or why not?

(iii) (4 Marks) Discuss an alternative regression approach for this type of analysis, mentioning any disadvantages in using this alternative approach.

(iv) (4 Marks) The Bland-Altman plot is a graphical technique commonly used for comparing methods of measurement. Explain how to construct and interpret this plot.

(b) The gold content of a concentrated sea-water sample was determined by using atomic-absorption spectrometry with the method of standard additions.

The results obtained were as follows:

```
>Gold <-c(30,40,50,60,70,0,10,20,80,70)
>Abso <-c(0.415,0.472,0.528,0.579,0.641,
 0.271,0.323,0.369,0.678,0.752)
>
> summary(lm(Abso~Gold))

Call:
lm(formula = Abso ~ Gold)

Residuals:
Min        1Q    Median        3Q       Max
-0.034662 -0.014833 -0.013924  0.004695  0.096057

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.2589060  0.0234791   11.03 4.07e-06 ***
Gold        0.0056720  0.0004668   12.15 1.95e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.03852 on 8 degrees of freedom
Multiple R-squared:  0.9486,    Adjusted R-squared:  0.9422
F-statistic: 147.6 on 1 and 8 DF,  p-value: 1.949e-06
```

(i) (5 Marks) Explain what the method of standard additions is, what it would be used to determine, and how regression analysis can be used as part of this analysis. Support your answer with sketches. **N.B.** You are not required to perform any calculations for this example

(c) In certain circumstances, Robust Regression may be used in preference to Ordinary Least Squares (OLS) Regression when fitting regression models. Answer the following questions relating to Robust Regression.

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

(ii) (1 Mark) State one difference between the OLS and Robust regression techniques in terms of computing regression equations.
(This question is continued on the next page.)

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

(iv) (2 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
18	22.91

Question 3. (25 marks) Experimental Design

(a) Specimens of milk from dairies in four different districts are assayed for their concentrations of the radioactive isotope Strontium-90. The results, in picocuries per litre, are shown in the table below.

District									Mean $\bar{x}_i$	St. Dev s_i
A	28.2	30.8	27.8	32.7	29.6	31.3	32.4	32.8	30.7	1.855
B	30.3	28.5	32.4	31.6	28.6	34.6	31.6	31.2	31.1	1.874
C	32.9	30.6	33.6	37.1	32.6	36.1	35.5	34.4	34.1	1.976
D	32.8	34.8	34	31.6	34.8	35.2	36.4	39.6	34.9	2.251
Overall									32.7	2.660

	DF	SS	MS	F
Between	?	?	?	?
Within	?	?	?	
Total	?			

The following `R` output has been produced as a result of analysis of these data:

(i) (8 Marks) Complete the ANOVA table in your answer sheet, replacing the "?" entries with the correct values.

(ii) (2 Marks) What hypothesis is being considered by this procedure.

(iii) (2 Marks) What is the conclusion following from the above analysis? State the null and alternative hypothesis clearly.

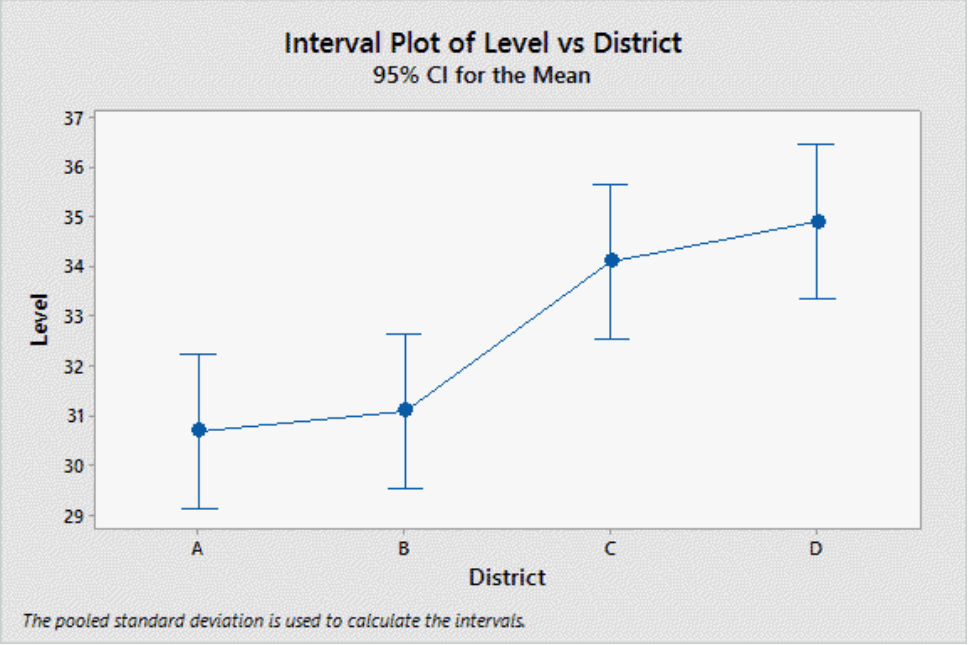
(b) (4 Marks)

The following outputs are Post-Hoc Procedures for the example in part (a). Interpret these outputs.

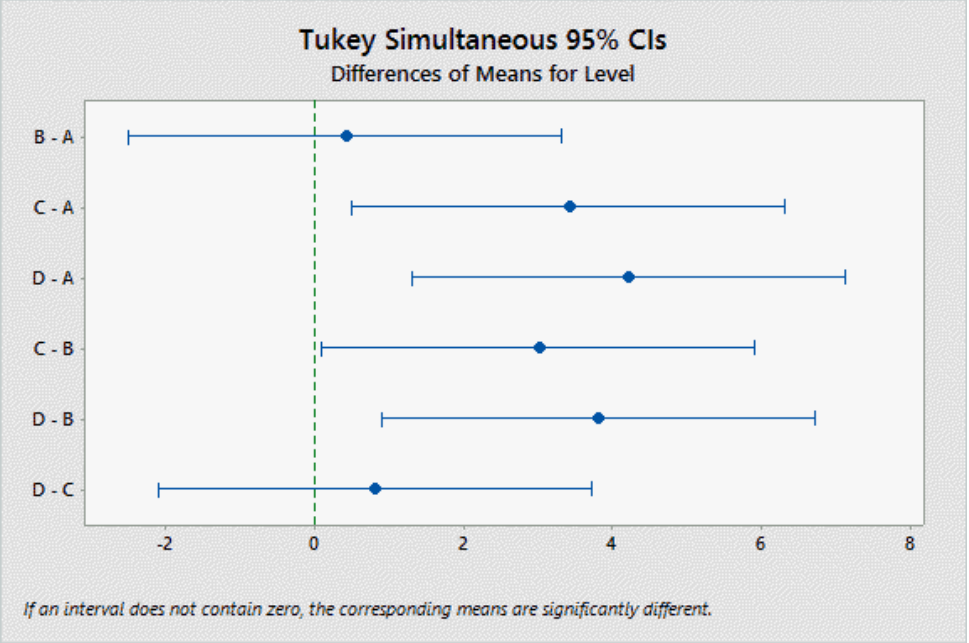
```
Tukey Pairwise Comparisons

Grouping Information Using the Tukey Method and 95% Confidence

District  N    Mean  Grouping
D         8  34.900  A
C         8  34.100  A
B         8  31.100    B
A         8  30.700    B
```



(The computer output continues on the next page.)



(c) Five standard solutions of chloride were prepared. Four titration methods, each with a different technique of end-point determination, were used to analyze each standard solution. The order of the experiments was randomized. The results of chloride found are shown below.

	Method U	Method V	Method W	Method X
Solution 1	10.15	10.64	10.71	10.71
Solution 2	10.13	10.23	10.72	10.01
Solution 3	10.32	10.22	10.60	10.24
Solution 4	10.29	10.72	10.53	10.58
Solution 5	9.81	10.28	10.20	9.99

You are also given the following information:

- $S_r^2 = 0.0394$
- $S_c^2 = 0.0305$
- The variance of concentrations in the table above is $\text{Var}(y) = 0.0771$

The following questions will result in the completion of the ANOVA Table on the next page. The $p-$values for both tests are already provided.

(i) (4 Marks) Complete the Sum of Squares column. (Show your workings.)

(ii) (1 Mark) State the degrees of freedom for the ANOVA Table.
medskip

(This question continues on the next page.)

(iii) (2 Marks) Based on the p-values, provided, what is your conclusion? Clearly state the null and alternative hypotheses for both tests.

Source	DF	SS	MS	F	p-value
Factor A	?	?	?	?	0.0130 *
Factor B	?	?	?	?	0.0196 *
Error	?	?	?		
Total	?	?			

Question 4. (25 Marks) Experimental Design

(a) (i) (3 Marks) What are the key components that need to be identified when designing an experiment?

(ii) (2 Marks) What is a randomised block design?

(iii) (2 Marks) What is an "a x b" factorial experimental design?

(iv) (2 Marks) What distinguishes a factorial experiment from a completely randomised experiment or a randomised block experiment?

(b) An experiment is run on an operating chemical process in which the aim is to reduce the amount of impurity produced. Three continuous variables are thought to affect impurity, these are agitation speed, concentration of NaOH and temperature. As an initial investigation two settings are selected for each variable these are

Factor:	low level	highlevel
Agitation speed (rpm)	15	30
Concentration of NaOH	40%	50%
Temperature (°F)	150	200

Readings were recorded of the impurity produced from the chemical process for each combination of the levels of these factors, and each combination was tested twice.

Agitation A	Conc NaOH C	Temperature T	Impurity Replicate 1	Impurity Replicate 2
-1	-1	-1	45.1	44.6
1	-1	-1	44.9	45.3
-1	1	-1	44.8	46.7
1	1	-1	44.7	44.8
-1	-1	1	33.0	35.0
1	-1	1	53.8	51.7
-1	1	1	32.6	33.7
1	1	1	54.2	53.2

(This question continues on the next page.)

The data was analysed with statistical computing software, creating the output presented below. Some numbers have been removed.

```
> summary(aov(y~A*C*T, Fact))

Df  Sum Sq Mean Sq  F value    Pr(>F)
A            1   .....   .....  405.762  3.85e-08 ***
C            1     0.1     0.1    0.115    0.7429
T            1   .....   .....   12.812    0.0072 **
A:C          1     0.1     0.1    0.083    0.7811
A:T          1   .....   .....  437.953  2.85e-08 ***
C:T          1     0.1     0.1    0.055    0.8200
A:C:T        1     2.3     2.3    2.540    0.1497
Residuals    8     7.3     0.9
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(i) (3 Marks) Calculate the contrasts, the effects and the sum of squares for the main effect A.

(ii) (3 Marks) Calculate the contrasts, the effects and the sum of squares for the interaction effect between A and T.

(iii) (3 Marks) Calculate the contrasts, the effects and the sum of squares for the interaction effect between A, C and T.

(iv) (3 Marks) Comment on the tests for significant for the main effects and interactions. State your conclusions clearly.

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

Question 5. (25 marks) Statistical Process Control

(a) Answer the following questions on graphical procedures used in Statistical Process Control.

Describe the purpose, the construction and interpretation of each chart or plot. Support your answers with sketches.

(i) (3 Marks) The CUSUM chart.

(ii) (3 Marks) The OC chart.

(This question continues on the next page.)

(b) (3 Marks) Write a brief description of the Mahalanobis Distance. Illustrate your answer with a sketch.

(c) 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 are 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 shown below.

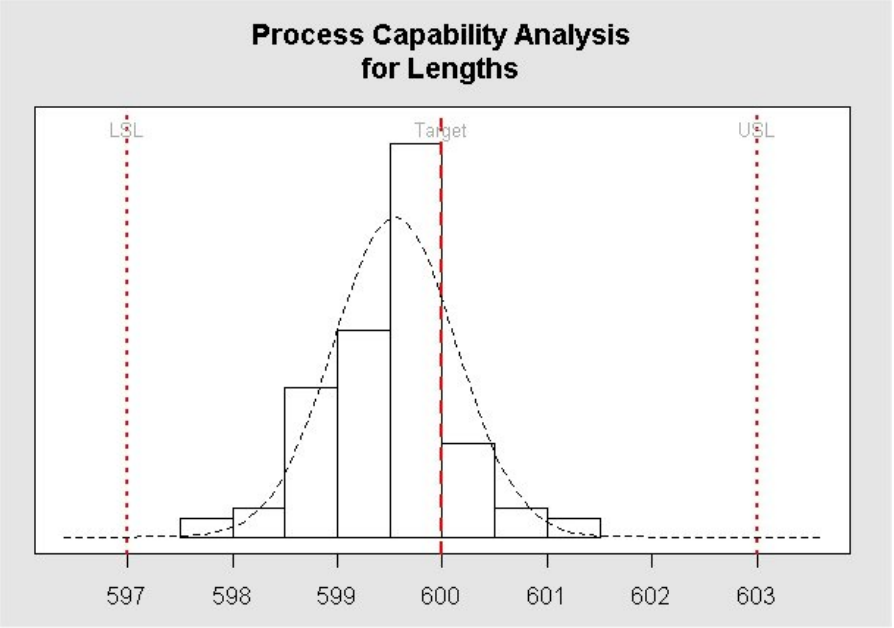
(ii) (2 Mark) Explain why there would be a discrepancy between C_p and C_{pk}. Illustrate your answer with sketches.

(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%
```



(d) (3×3 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.

Discuss any three of these rules, stating their mathematical basis 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 X.

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

Formulas and Tables

Critical Values for Dixon Q Test

N	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
3	0.941	0.97	0.994
4	0.765	0.829	0.926
5	0.642	0.71	0.821
6	0.56	0.625	0.74
7	0.507	0.568	0.68
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	0.41	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

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

Control Limits for Control Charts

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

$$\bar{\bar{x}} \pm 3\frac{\bar{s}}{c_4\sqrt{n}} \qquad \left[\bar{R}D_3, \bar{R}D_4\right]$$

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

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

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

Process Capability Indices

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

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

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