CHAPTER 1

INTRODUCTION

1.1 Background of Study

A heavy metal is a member of an ill-defined subset of elements that exhibit metallic properties, which would mainly include the transition metals, some metalloids, lanthanides, and actinides. Heavy metal can include elements lighter than carbon and can exclude some of the heaviest metals. Heavy metals occur naturally in the ecosystem with large variations in concentration. In modern times, anthropogenic sources of heavy metals, i.e. pollution, have been introduced to the ecosystem. Waste-derived fuels are especially prone to contain heavy metals so they should be a central concern in a consideration of their use. Living organisms require varying amounts of "heavy metals." Iron, cobalt, copper, manganese, molybdenum, and zinc are required by humans. Excessive levels can be damaging to the organism. Other heavy metals such as mercury, plutonium, and lead are toxic metals that have no known vital or beneficial effect on organisms, and their accumulation over time in the bodies of animals can cause serious illness. Motivations for controlling heavy metal concentrations in gas streams are diverse. Some of them are dangerous to health or to the environment (e.g. mercury, cadmium, arsenic, lead, chromium), some may cause corrosion (e.g. zinc, lead), some are harmful in other ways (e.g. arsenic may pollute catalysts). Within the European community the thirteen elements of highest concern are arsenic, cadmium, cobalt, chromium, copper, mercury, manganese, nickel, lead, tin, and thallium, the emissions of which are regulated in waste incinerators. Some of these elements are actually are carcinogenic or toxic, affecting, among others, the central nervous system (manganese, mercury, lead, arsenic), the kidneys or liver (mercury, lead, cadmium, copper) or skin, bones, or teeth (nickel, cadmium, copper, chromium).

Heavy metal pollution can arise from many sources but most commonly arises from the purification of metals such as the smelting of copper and the preparation of nuclear fuels. Through precipitation of their compounds or by ion exchange into soils and mud, heavy metal pollutants can localize and lay dormant. Unlike organic pollutants, heavy metals do not decay and thus pose a different kind of challenge for remediation. Scientists use many methods to identify the heavy metals, the easiest and widely use method is UV-VIS spectrum analysis where the samples are test for it ion concentration using spectrometer.

A spectrometer (spectrophotometer, spectrograph or spectroscope) is an instrument used to measure properties of light over a specific portion of the electromagnetic spectrum, typically used in spectroscopic analysis to identify materials. The variable measured is most often the light's intensity but could also, for instance, be the polarization state. The independent variable is usually the wavelength of the light or a unit directly proportional to the photon energy, such as wave number or electron volts, which has a reciprocal relationship to wavelength. A spectrometer is used in spectroscopy for producing spectral lines and measuring their wavelengths and intensities. The intensities difference (absorbtivity) of the test sample then is use to predict it concentration by referring to standard calibration curve for that specified metals ions.

Calibration curve is a general method for determining the concentration of a substance in an unknown sample by comparing the unknown to a set of standard samples of known concentration. In this project these concentration values will be measured and correlate to their absorptivity values which are analyzed using spectrometer. Then this data is used to construct the calibration curve graph. This graph will represent the behavioral change of absorptivity of that heavy metals to it concentration. By using the mathematical relationship derives from it, unknown concentration sample can be estimated given known absorptivity value of the sample. The concentrations of the standards must lie within the working range .Conventionally; the calibration curve with linear relationship is preferred which constructed using single point reference method.

1.2 Problem Statement

In UV-VIS spectrum analysis, a calibration curve is a general method for determining the concentration of a substance in an unknown test sample by comparing the unknown value to a set of standard samples of known value.[1] . The calibration curve for spectrometer is a plot of how the heavy metals ion response, in term of it absorptivity, changes with the concentration of the analyte (the substance to be measured). The operator prepares a series of standards across a range of concentrations of known concentration and test it in spectrometer for it correspond absorptivity for each of the concentration. For most analyses a plot of instrument response vs. analyte concentration will show a near linear relationship. The operator can measure the response of the unknown and, using the calibration curve, can interpolate to find the concentration of analyte for that unknown sample. In this case, concentration of heavy metal ion is predicted by referring it absorptivity value to the calibration curve.

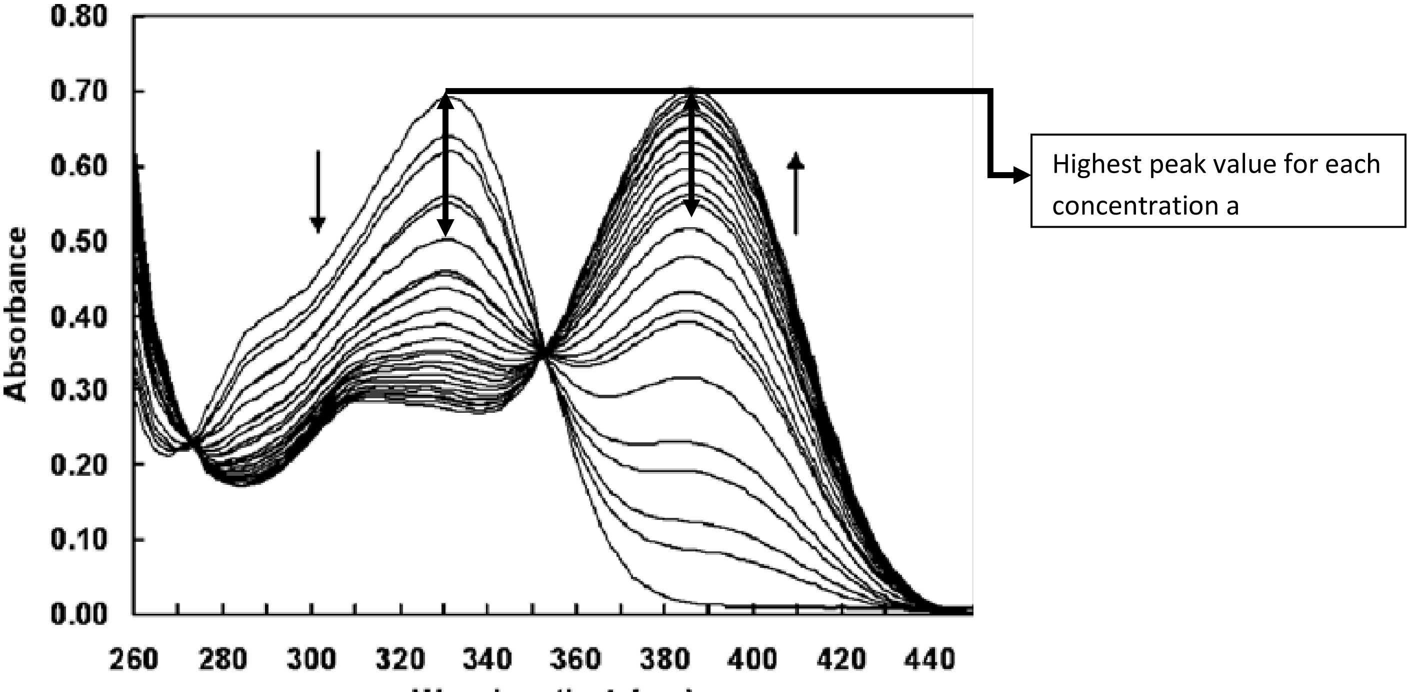


Figure1.0: UV–Vis absorption changes during the formation of heavy metal

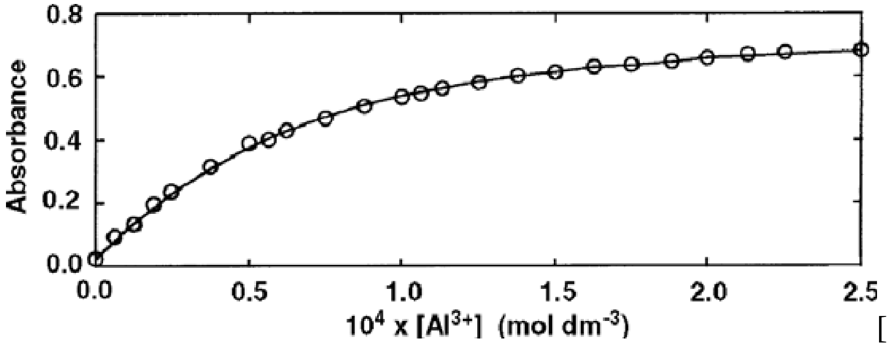
Conventionally, the method use is single point sampling with linear model fitting. In many cases, the correlation between predictor (absorbance intensity) and respondent (test sample concentration) is non linear. The project is to improve the accuracy of the calibration construction method to provide more reliable result in testing heavy metals using UV-Vis analysis.

a) Single point method

1. Take highest peak from the absorbance versus wavelength graph as the absorptivity value for correspond concentration.
2. Plot the absorbance versus concentration as calibration curve.
3. Transform the graph to function model.

b) Multi point method

1. Take value at several locations on graph as the absorptivity value for correspond concentration.
2. Represent all the point as one point for each concentration by using Principle Component Regression Analysis, or Partial Least Square Regression.
3. Plot the graph between predictor variable (score or intensity) to it corresponding concentration.
4. Transform the graph to function model.



[1]

Figure 1.1: Relationship between absorbance and concentration

Different between points is the data extraction method, where instead of taking the value of absorbance intensity, we use the different between the intensity as a predictor variable. The application of this technique can also be use in both PCA and PLSR.

Principal component analysis (PCA) creates new variables (components) that consist of uncorrelated, linear combinations of the original variables. It is used to simplify the data structure and still account for as much of the total variation in the original data as possible. By using PCA, we try to reduce numbers of variables to a score which represent most of the original data which can be use as a single variable.

Partial least square regression (PLSR) function almost same as PCA but in addition, it include the calculation to get better regression of graph during the transforming of many variable to single variable process.

1.3 Objective

1) To evaluate the performance of a linear-single point calibration method (base approach). - -

   - Since the modifications are compared to base approach, evaluation concerning this method must be thoroughly made. Efficiency tests of original calibration curve construction method are done both in term of fitting accuracy and it reliability to estimate unknown heavy metal sample concentration based on it absorptivity.

2) To evaluate nonlinear correlations to improve calibration accuracy.

   - Since the linear relationship shown low accuracy, study will be conducted to test other alternative of correlation between variable apart from linear relationship. Nonlinear correlations are introduced into the procedure until the best fit relationship is found.

3) To identify important factor in improving the accuracy of calibration curve

   - Modification is done in each step of calibration curve construction. Every alternative will be experimented including best point extraction location and quantities for both single point and multiple point calibration curve, the preprocessing of raw data and the transformation of preprocess data with multivariate analysis.

4) To compare accuracy of modified procedure calibration curve to a base approach.

- Each modification are tested in term of it accuracy and the result are compared to base approach. Any improvement is noted and further modification considers the newly found improvement as part of it procedure. Comparison is done in term of percentage error different compared to base approach.

1.4 Scope of Study

In order to achieve the objective outlined, the project's study will cover several alternative modifications. The experiment will be conducted to each and every branch of calibration curve construction step as above:

❑ Calibration curve models

- Linear (base approach)
- Non-Linear (Quadratic, Gaussian, Cubic)

❑ Single and multiple point extraction

- Point location
- Number of point
- Methods of extraction (preprocessing efficiency)

❑ Multivariate analysis for processing multiple points

- PCA
- PLRS

1.5 The Relevancy of the Project

- Data from this project can be use to improve the process analysis of concentration
- Verifying the proper functioning of an analytical instrument or a sensor device such as an ion selective electrode
- The improved calibration curve might be use to determine the basic effects of a control treatment (such as a dose-survival curve in clonogenic assay)
- Factory can use same calibration curve technique to calibrate their machine setting to get desired specification.

1.6 Feasibility of the Project within the Scope and Time frame

During first semester, student needs to focus on literature review part of the project. One must have deep understanding about the procedures, steps and calculation necessary to conduct the project. Student will also need to focus on learning and experimenting with the method use in this project. One must be able to utilize Mat Lab software fully in order to ease the experimenting process. The final target of first semester will be at least to conduct one experiment and understand advantage and disadvantages of this new method and get an insight on how to improve it to meet expected result.

During second semester, many experiment need to be conduct. The purpose is trying to find best possible combination of location, interpretation of sample and visual representation of the calibration curve. During this semester, student should able to compare the efficiency of each method and find best calibration construction technique.

CHAPTER 2

THEORY AND LITERATURE REVIEW

2.1 Spectrometer

Spectrometer is a tool use to measure wavelengths or indexes of refraction. Spectroscopy is a simple and powerful method for performing both qualitative and quantitative analyses. Each chemical species has a unique spectral fingerprint based on where electrons are located with respect to the nucleus.

Chemists commonly use absorbance spectroscopy, or how a substance absorbs photons of light, to obtain both qualitative (identity) and quantitative (amount) information. The quantitative measurement is achieved because each photon of light absorbed corresponds to the excitation of a single electron.

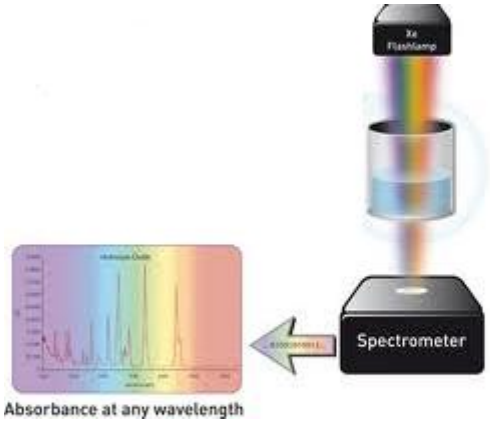


Figure 2.0: The spectrometer machine

2.2 UV-Vis Spectrum Analysis

Beer–Lambert law: Lambert law states that there is a logarithmic dependence between the transmission (aka transmissivity), *T*, of light through a substance and the product of the absorption coefficient of the substance, *α*, and the distance the light travels through the material (i.e. the path length), *ℓ*. The absorption coefficient can, in turn, be written as a product of either a molar absorptivity of the absorber, *ε*, and the concentration *c* of absorbing species in the material, or an absorption cross section, *σ*, and the (number) density *N* of absorbers

$$T = \frac{I}{I_0} = 10^{-\alpha\ell} = 10^{-\varepsilon\ell c}$$

The transmission (or transmissivity) is expressed in terms of an absorbance which for liquids is defined as

$$A = -\log_{10}\left(\frac{I}{I_0}\right)$$

This implies that the_absorbance becomes linear with the concentration (or number density of absorbers) according to:

$$A = \varepsilon\ell c = \alpha\ell$$

Thus, if the path length and the molar absorptivity or the absorption cross section is known and the absorbance is measured, the concentration of the substance (or the number density of absorbers) can be deduced.

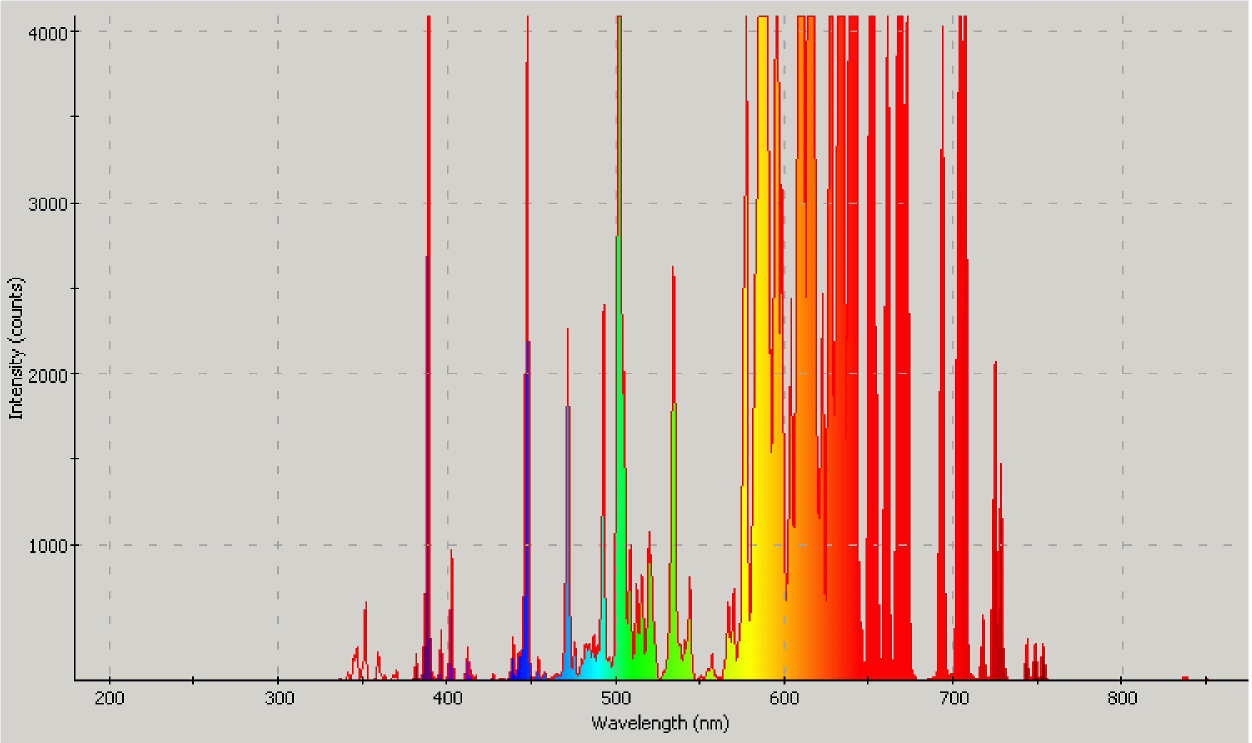


Figure 2.1: The graph wavelength versus intensity produced by spectrometer.

2.3 Principle Component Analysis

In order to detect heavy metals in environment, we need first to know molar absorptivity constant for it. By controlling the concentration of metal ions in test sample and it path length, we try to find it molar absorptivity constant. However after several experiments, the result shows logarithmic relationship instead linear behavior between the concentration of the metal ion and it absorbance. This gives us unstable molar absorptivity constant.

In order to fix this problem, we came up with principle Component Analysis method (PCA). The purpose is to reduce the dimensionality of a data set (sample) by finding a new set of variables, smaller than the original set of variables, which nonetheless retains most of the sample's information. It involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components.

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. PCA is theoretically the optimum transform for given data in least square terms.

For a data matrix, X^T, with zero empirical mean (the empirical mean of the distribution has been subtracted from the data set), where each row represents a different repetition of the experiment, and each column gives the results from a particular probe, the PCA transformation[2] is given by:

$$\begin{aligned}\mathbf{Y}^{\mathrm{T}} &= \mathbf{X}^{\mathrm{T}}\mathbf{W} \\ &= \mathbf{V}\mathbf{\Sigma}^{\mathrm{T}}\end{aligned}$$

where the matrix Σ is an *m*-by-*n* diagonal matrix with nonnegative real numbers on the diagonal and $W \Sigma V^T$ is the singular value decomposition (svd) of X.

Given a set of points in Euclidean space, the first principal component (the eigenvector with the largest eigenvalue) corresponds to a line that passes through the mean and minimizes sum squared error with those points. The second principal component corresponds to the same concept after all correlation with the first principal component has been subtracted out from the points.

Each eigenvalue indicates the portion of the variance that is correlated with each eigenvector. Thus, the sum of all the eigenvalues is equal to the sum squared distance of the points with their mean divided by the number of dimensions. PCA essentially rotates the set of points around their mean in order to align with the first few principal components. This moves as much of the variance as possible (using a linear transformation) into the first few dimensions. The values in the remaining dimensions, therefore, tend to be highly correlated and may be dropped with minimal loss of information. PCA is often used in this manner for dimensionality reduction. PCA has the distinction of being the optimal linear transformation for keeping the subspace that has largest variance.

2.4 Partial Least Square Regression

Partial least squares regression is an extension of the multiple linear regression model. In its simplest form, a linear model specifies the (linear) relationship between a dependent (response) variable *Y*, and a set of predictor variables, the *X*'s, so that

$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p$

In this equation b_0 is the regression coefficient for the intercept and the b_i values are the regression coefficients (for variables 1 through *p*) computed from the data.

So for example, you could estimate (i.e., predict) a person's weight as a function of the person's height and gender. You could use linear regression to estimate the respective regression coefficients from a sample of data, measuring height, weight, and observing the subjects' gender. For many data analysis problems, estimates of the linear relationships between variables are adequate to describe the observed data, and to make reasonable predictions for new observations.

The multiple linear regression model has been extended in a number of ways to address more sophisticated data analysis problems. The multiple linear regression model serves as the basis for a number of multivariate methods such as discriminant analysis (i.e., the prediction of group membership from the levels of continuous predictor variables), principal components regression (i.e., the prediction of responses on the dependent variables from factors underlying the levels of the predictor variables), and canonical correlation (i.e., the prediction of factors underlying responses on the dependent variables from factors underlying the levels of the predictor variables). These multivariate methods all have two important properties in common. These methods impose restrictions such that (1) factors underlying the *Y* and *X* variables are extracted from the *Y'Y* and *X'X* matrices, respectively, and never from cross-product matrices involving both the *Y* and *X* variables, and (2) the number of prediction functions can never exceed the minimum of the number of *Y* variables and *X* variables.

Partial least squares regression extends multiple linear regression without imposing the restrictions employed by discriminant analysis, principal components regression, and canonical correlation. In partial least squares regression, prediction functions are represented by factors extracted from the *Y'XX'Y* matrix. The number of such prediction functions that can be extracted typically will exceed the maximum of the number of *Y* and *X* variables.

In short, partial least squares regression is probably the least restrictive of the various multivariate extensions of the multiple linear regression model. This flexibility allows it to be used in situations where the use of traditional multivariate methods is severely limited, such as when there are fewer observations than predictor variables. Furthermore, partial least squares regression can be used as an exploratory analysis tool to select suitable predictor variables and to identify outliers before classical linear regression.

The general underlying model of multivariate PLS is:

$$X = TP^{\top} + E$$
$$Y = TQ^{\top} + F,$$

where *X* is an $n \times m$matrix of predictors, *Y* is an $n \times p$matrix of responses, *T* is an $n \times l$ matrix (the *score*, *component* or *factor* matrix), *P* and *Q* are, respectively, $m \times l$and $p \times l$ *loading* matrices, and matrices *E* and *F* are the error terms, assumed to be i.i.d. normal.

A number of variants of PLS exist for estimating the factor and loading matrices *T*,*P* and *Q*. Most of them construct estimates of the linear regression between *X* and *Y* in:

$$Y = X\tilde{B} + \tilde{B}_0.$$

Some PLS algorithms are only appropriate for the case where *Y* is a column vector, while others deal with the general case of a matrix *Y*. Algorithms also differ on whether they estimate the factor matrix *T* as an orthogonal, an orthonormal matrix or not. The final prediction will be the same for all these varieties of PLS, but the components will differ.

Partial least squares regression has been used in various disciplines such as chemistry, economics, medicine, psychology, and pharmaceutical science where predictive linear modeling, especially with a large number of predictors, is necessary. Especially in chemometrics, partial least squares regression has become a standard tool for modeling linear relations between multivariate measurements (de Jong, 1993).

CHAPTER 3

METHODOLOGY AND PROJECT WORK

General methodology

In making the calibration curve, each modification is done a bit by bit. Earlier step are experimented with every alternative and tested for it accuracy. Procedure with better accuracy are choose and be used in the next modification. The process will be repeated until we exhausted every possible alternative covered by scope of study. Although some of the procedure used is different, most of it can be explain using general methodology below:

1. Test the heavy metal ion samples of known concentration for it absorptivity.
2. Construct the calibration curve based on single point linear fit model (base case).
3. Construct the calibration curve based on:
    1. Different model fit (linear, quadratic, cubic , Gaussian)
    2. Different numbers of point extraction quantities (1-5)
    3. Different extraction point location (extremum)
    4. Different raw data manipulation method (different between Pt)
    5. Different data transformation method (PCR,PLSR)
4. Compare the accuracy of difference approach from base case.

Raw data extraction location

First step in constructing the calibration curve is the extraction of the heavy metal sample absorptivity's intensity. This absorptivity value which are considered raw data, are taken by selecting the location of suitable wavelength. Since the conventional method use highest peak as their extraction point, this project also adapt same technique for a same reasons. The visibility of highest peak can help the scientist to maintain uniform extraction location for all the concentration thus helping reduce human error in the construction process afterward.

Apart from single point usual location which is point 3(refer graph 3.0 below), the experiment to find better location point are done toward other extremum location including valley of graph. The extraction locations are chosen cover wider coverage of spectrum possible without noise.

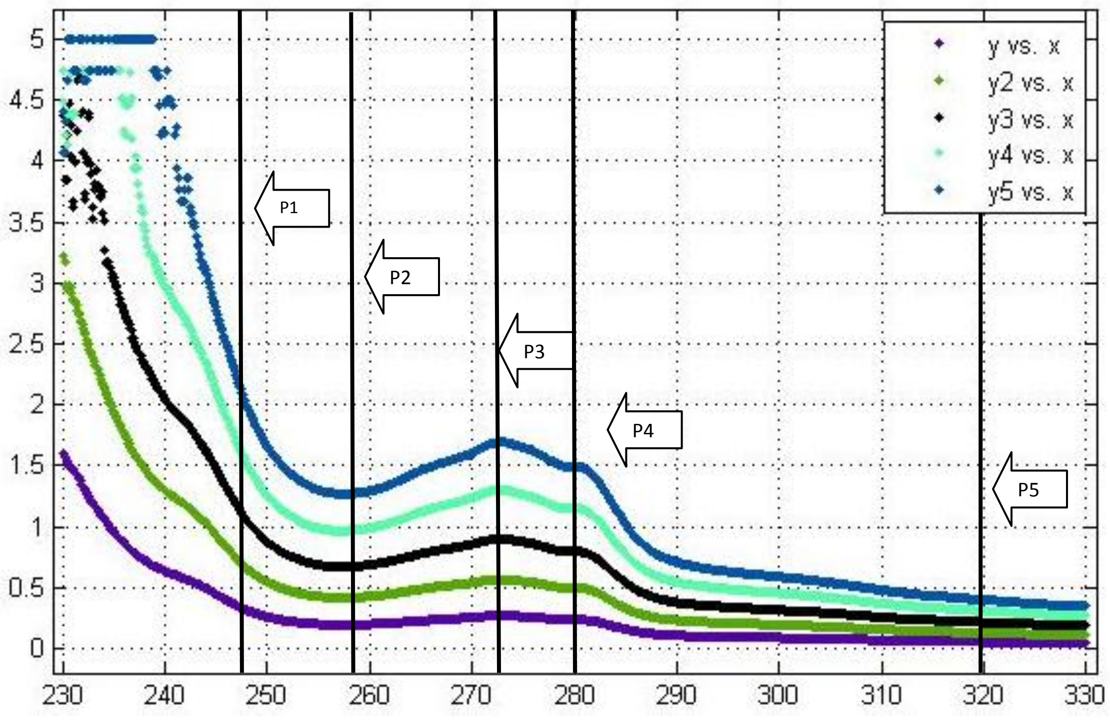


Figure 3.0: Graph raw data extraction location from spectrum

General calibration curve construction flowchart

Flow chart above show the basic outline of the calibration curve design procedure. Step 2 to 5 is varied based on data transformation method used. Basically, the modification is done step by step, in each step new construction alternative are done and best option (least error) are choose for next step modification.

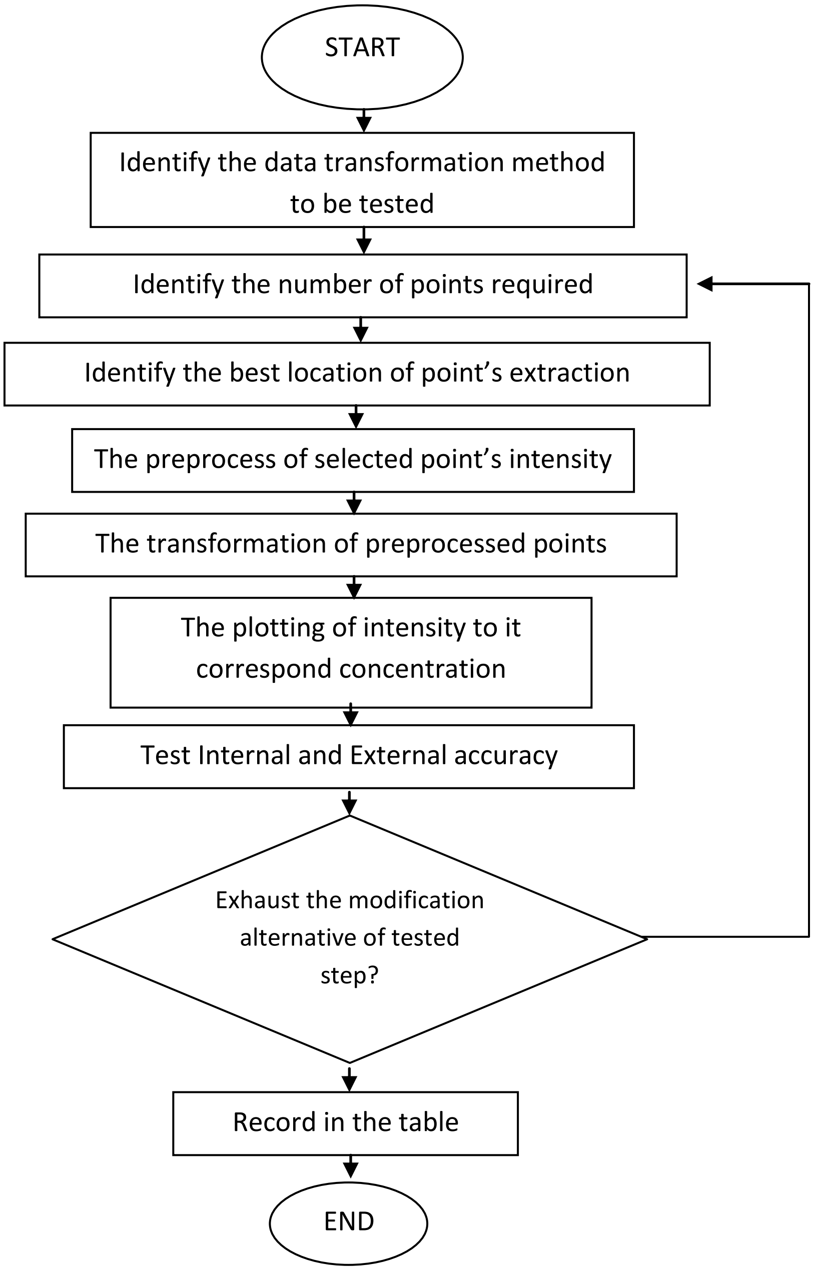

START
Identify the data transformation method to be tested
Identify the number of points required
Identify the best location of point's extraction
The preprocess of selected point's intensity
The transformation of preprocessed points
The plotting of intensity to it correspond concentration
Test Internal and External accuracy
Exhaust the modification alternative of tested step?
Record in the table
END

Figure 31: flowchart of calibration curve construction

Modification process ladder

Modification are done step by step. Each part of experiment deal with one aspect of the calibration curve design. The table below shows the approach taken to improve the construction procedure beginning with single point and then proceed to multiple point based design afterward. Single point method includes the testing to find the best peak location between point 1 and point 2 and then testing of the best model fit for calibration curve whether it is original linear relationship, quadratic relationship, cubic relationship or even Gaussian are considered.

Multiple point based construction method introduces new whole experiment. Raw data are preprocess in a way that instead taking it actual value (absorbance intensity), the predictor variable use the different between 5 selected point locations. Each of 6 point different that been used as raw data are evaluated and the lowest error is choose. After undergo preprocess, the selected points are transform to single variable using multivariate analysis. This part evaluate the accuracy of infusing the preprocess method and raw data at first, and then using raw data only for the next evaluation. The partial least square method is same in term of procedure apart from it considered the linear regression in it calculation. The experiment will show which among the two of multivariate analysis which are better method to be use for different requirement necessary.

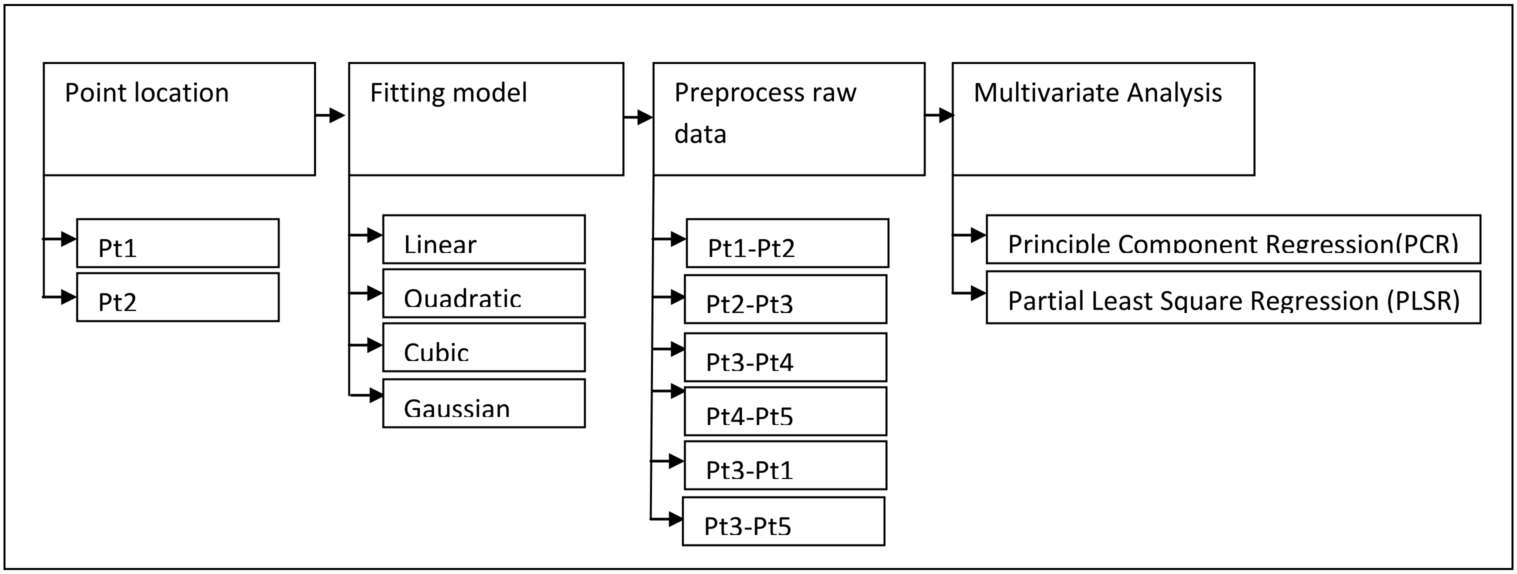


Figure 3.2: The modification ladder

Result

Case Study	Function Model	Point taken	Validation SSE	Fitted SSE	Accuracy improvement
Single point	Linear	Pt 3	5.214	3.529	-
	Quadratic	Pt 3	1.797	3.04	65.5
	Cubic	Pt 3	107.622	4.42e-029	-
	Gaussian	Pt 3	3.973	3.926	23.8
Single point	Linear	Pt 4	54.110	2.930	-
	Quadratic	Pt 4	134.777	2.373	-
	Cubic	Pt 4	5.971e+003	1.546	-
	Gaussian	Pt 4	15.371	3.191	-
Multiple points (difft. In Pt)	Linear	Pt1-Pt2	4.370	0.0412	16.2
	Linear	Pt2-Pt3	4.564	0.081	12.5
	Linear	Pt3-Pt4	3.756	0.043	28.0
	Linear	Pt4-Pt5	4.415	0.056	15.3
	Linear	Pt3-Pt1	4.407	0.052	15.5
	Linear	Pt3-Pt5	4.296	0.053	17.6
	Quadratic	Pt3-Pt4	1.688	3.052	67.6
Multiple points (PCR + difft. In Pt)	Linear	Pt3-Pt4 and Pt2	5.454	1.141	-
	Quadratic	Pt3-Pt4 and Pt2	2.163	2.628	58.5
Multiple points (PCR analysis)	Linear	Pt2,Pt3 and Pt4	6.053	1.432	-
	Quadratic	Pt2,Pt3 and Pt4	2.095	2.755	59.8
Multiple points (PLS + difft. In Pt)	Linear	Pt3-Pt4 and Pt2	2.448	0.120	53.0
	Quadratic	Pt3-Pt4 and Pt2	1.745	0.107	66.5
Multiple point(PLS analysis)	Linear	Pt2,Pt3 and Pt4	2.640	0.112	49.4
	Quadratic	Pt2,Pt3 and Pt4	1.958	0.135	62.5

Table 3.0: Accuracy comparison between base approach and modified approach

Discussion

Part 1

a) The important of fitting function model

From the result, we can say that the fitting model is major influence in improving the calibration curve. By changing the model representation alone, we able to reduce sum of square errors reduction from traditional method significantly. Since the procedure is simple, the improvement considered feasible for application for most situation.

b) Best fitting model for single point based calibration curve

From the graph, we understand that quadratic based fitting model is the highest in accuracy to represent the relationship between concentration and the absorbance. By changing it alone, we can reduce SSE (sum of square error) for future estimation for both interpolation and extrapolation estimation up to 65.5%. The next best fitting model is Gaussian with 23.8% reduction in SSE followed by liner model. Meanwhile, cubic is considered unpractical of usage since it SSE far worst compared to the original model.

Part 2

a) Important of using different in point compared to actual value

From the result, we can see that External SSE or Validation of calibration curve is within 4.56 to 3.76 for all linear based models. This alone proven that by taking different between points, the error is lower compared to directly using the intensity value as the predictor variable. It further proven by Internal SSE test where every location posed lower value thus lower distant between graph line and point variable (curve fitting). Since this method produce higher accuracy, further modification of the calibration curve will be done using this way.

b) Important of point location selection

Table 2.2 shows that although the same extraction method (different between points) is used, the accuracy of calibration curve estimation varies between each other. This shows that different point has a potential to provide better predictor variable for our calibration curve.

c) Best location point location

From the table, different between point 3 and point 4 provide lowest SSE. The different between Pt3 (highest peak) and Pt4 (2nd peak) has an external SSE as low as 3.756 which is 28% lower compare to extraction from actual intensity. Without changing fitting model of calibration curve, we are able to identify this location as best location for data extraction.

d) Location selection (peak-extremum)

Since the purpose of this part experiment is to find the best location, the extremum value data are chosen for 3 main reasons:

I) Easiness

- Visibility- To ease the user, we choose the extremum for all 5 location, since it can be notice easier and avoid complication of finding it in future.

- Conventionally used- Since the peak are conventionally use, the method are more user friendly since it also consider peak as main location.

- Simplicity in calculation – if this method are proven to be effective, peak value intensity are easy to read hence ease the calculation procedure.

II) Accuracy

- Free from interference/noise – since the location far from minimum and maximum wavelength, the interference/noise are kept at lowest possible.

- Uniform increment/ interval – by choosing location that have steady/comparable increment between each concentration, chance of higher accuracy of calibration curve are also improve.

III) Method based

- Suitability in calculation –each location are picked based on it suitability of calibration curve construction method that will be applied to the data.

Part 3

a) The potential of different between point

By using new model fit, we are able to reduce the External SSE considerably. This phenomenon reveal that by changing model fit, calibration curve with using different between point can be further improve as same as when using real intensity value(original extraction method). This indirectly shows that this extraction method probably can be used in future instead of original intensity. This experiment is able to reduce the error up to 67.6% which is the highest so far.

Part 4

a) Effect of PCA in procedures

From the result, we can say that PCA (principle component) analysis pose no improvement compare to previous method. Although we transform fitting model to quadratic function fit, it still fall lower compare to previous method. Nevertheless, experiment shows that this method can reduce the error of normal calibration curve up to59.8% and 58.5% for both first part and second part of experiment respectively.

b) Methodology

By comparing both procedures, it seems the part4a are harder to conduct since it required additional step in finding the different point. Given that it are lower in accuracy making this method are less favorable compare to part3b. Overall, since the procedure for using PCA is more complicated and the accuracy improvement are lower compare to previous method which more simple, this method are considered unfeasible for calibration curve design application.

c) Probable reason of low accuracy

By using just two points as reference, the PCA can only get lowest percentage of overall data thus contributing low degree of accuracy. In next experiment, more quantity needs to be use until satisfactory accuracy reached. The location of the data derived also need to be varied to find better spot; this will also require the further experiment to test the validity of selected location.

Part 5

a) Effect of PLSR in procedures

From the result, we can say that PLSR (partial least square regression) didn’t improve much the result in quadratic model of calibration curve, however this method able to produce highest accuracy for linear based fitting model. Without transforming to quadratic model, PLSR can improve the accuracy up to 53%.Overall, PLSR method are better suited for calibration curve of spectrometer compare to PCA since both Internal SSE and External SSE of this method are better compare to PCA method.

b) Probable reason of low accuracy for quadratic fitting

Since PLSR method already considers the regression when calculating the score, it was able to produce best linear fitting straight away. In contrast, for the same reason it accuracy can’t be improved much when changing to quadratic fitting.

General Discussion

a) The External SSE

External SSE use comparison of 2 omitted data with calculated concentration based on calibration curve model constructed. The model’s accuracy was calculated using sum of square error method. Both the omitted concentration cover interpolation and extrapolation. Since the omitted data have not involved anyway in construction process, the reliability of this method accuracy test are kept close to real problem as possible.

b) The Internal SSE

Internal SSE use the same accuracy method with External SSE where sum of square different between actual data and the calculated data based on newly constructed calibration curve are used. The different is, the calculated data used were not are the predicted one instead the data are what were used in construction procedure. The comparison simply between the calibration graph line to the graph point at same concentration. It represent the line fitting of calibration curve.

c) Internal SSE Vs external SSE

External data are prioritize over internal because the model itself are construct based on internal/raw data thus it cannot predict future concentration which is outside of it. External SSE on the other hand provides estimation as good as any other potential future problem since although it originally comes from the same raw data; it takes no part whatsoever in the calibration curve construction.

d) Overall comparison

The results show that the relationship between predictor variable (absorbance intensity) and respond variable (concentration) are highly non-liner. Thus by using original method which is linear fitting model, the calibration curves tend to produce highest error. By changing the model to quadratic, the accuracy can be improved up to 65%. Although cubic produce lower Internal SSE (fitting error) the accuracy are still low compare to quadratic (overfitting phenomena). Apart from that, we understand that by using different between points instead of point's intensity, we are able to increase the accuracy higher. The results also reveal that generally multiple point based are better compare to single point based calibration curve. Compare to PCA, PLSR method yield highest accuracy both in quadratic and linear model fitting. Overall, the best quadratic model is using different between point method (Pt3 and Pt4) and the best linear model is PLSR method of point 2 and different between point (Pt3 and Pt4).

3.3 Conclusion

Throughout the whole project, each and every procedure modification and accuracy testing manage to give us the overview of the feasibility of each modification and it effect to the calibration curve accuracy. The study were able fulfilled the objective in term of testing the best possible calibration curve construction procedure with exploring every alternative covered by scope of study as intended.

Experiment show that single point based linear model fitting is not the best calibration curve criteria. It is higher in errors both in term of model fitting and it ability to estimate the sample's concentration based on it absorptivity value. Based on model fitting, linear relationship that it used are not suitable since the actual data variables are behaving in non linear function. Since the calibration curve just use single point from the spectrum, the graph didn't have enough data to predict future concentration of heavy metal sample.

Experiment show that quadratic model is the best data representative to explain the relationship between heavy metal's absorptivity value and it corresponding concentration. By changing from linear to quadratic model, errors can be reduced significantly. This statement hold true almost in all of the method and the modification made except for the PLSR. That method alone favors linear regression over quadratic model since the improvement in accuracy very small and deemed infeasible.

Experiment show that there are several factors which is important in determining better calibration curve apart from model fitting. The detail of location which is consisting of location and quantity of point extracted from spectrum are able to decide whether the curve designed are improved or worst than base approach. Study show best extraction location is the highest peak without noise follow by second highest peak for single point.   For multiple point method, the different between points two point locations mentioned above are considered the best option. Preprocessing step for multiple point method is proven produce higher accuracy compare when the raw data are use as predictor variable. Among two multivariate analyses, PLSR analysis used in the procedure can improved an accuracy further compare to PCR analysis.

Experiment show that the modification in procedure is able to improve the original method significantly. By comparison, the best calibration curve for linear relationship is when using PLSR with the preprocessing step of point 3 and point 4 with point 2. For quadratic, it suffices just by changing the relationship from linear to quadratic model and using preprocessing without multivariate. Overall, by modification of calibration curve construction procedure, it error can be remove almost up to 70%. Since the modifications are fairly simple, it is considered feasible and relevant to be applied into current technique.

Project Gantt chart

No.	Detail/ Week	1	2	3	4	5	6	7		8	9	10	11	12	13	14	15
1	Project Work Continues	🟪	🟪	🟪	🟪	🟪	🟪	🟪	Mid-Semester Break								
2	Submission of Progress Report									●							
3	Project Work Continues									🟪	🟪	🟪	🟪	🟪			
4	Pre-EDX												●				
5	Submission of Draft Report													●			
6	Submission of Dissertation (soft bound)														●		
7	Submission of Technical Paper														●		
8	Oral Presentation															●	
9	Submission of Project Dissertation (Hard Bound)																●

● Suggested milestone

🟪 Process