Introduction

Vapor Liquid Equilibrium has long been an important model for industrial use in many techniques, such as distillation and separation process. In contemporary world, many of these VLE models are complex, non-linear model which are solved with powerful mathematical program with high accuracy and precision. However, high specifications computers are required for such task to reduce waiting time, which are not ubiquitous for access. Luckily, most commons component equilibrium model has been collected in database and can be found online. However, more specialized components and complex mixture could be unavailable and may be complicated to identify the correct activity model. Coupled with idle time of low performance computer, this could be disastrous for real-time research and development. Fortunately, statistical analysis and methods could be used to manipulate non-linear activity model into linear form and solved with linear statistical model. This paper is aimed at exploring the statistical application of the Maximum Likelihood Estimation (MLE) and Multivariate Linear Model (MLM) onto estimating activity coefficient of one-parameter and two-parameter Margules model in Vapor-Liquid Equilibrium problems.

Overall, the activity coefficient of one and two-parameter Margules equation estimated using the two statistical models are precise and accurate for linear data trends. However, when facing data exhibiting non-linear trends, these models failed to estimate parameter accurately. When face with random errors, MLE method in general is preferred for estimation, as it is more precise than MLM in estimating activity coefficient for both one and two-parameter Margules equation, despite having a lower accuracy but not significant for two-parameter Margules estimation. Despite the linear model fails to accurately portray points with curvatures, it has been shown that different approaches of linear methods might be better than others. Further exploration of other linear models are encouraged as there may be better linear methods for estimation beside the MLM and MLE methods.

Methods

Data collection:

Two statistical technique of MLE and MLM are used to model a 2 VLE mixture: methanol (1) and ethanol (2) (1^{st} mixture), and isopropanol (1) and water (2) (2^{nd} mixture). P-x-y data of these 2 VLE mixture were extracted from VLE-calc.com with varying Pressure at constant T = 25 Celcius. Data is then being added with randomized generations to reflect real-time data collection error.

Activity Coefficient Model and Raoult's law (from Carl's):

One-Parameter activity coefficient model:

$ln\gamma_1 = A_{12}x_2^2$; $ln\gamma_2 = A_{12}x_1^2$;

Two-Parameter activity coefficient model:

$ln\gamma_1 = x_2^2[A_{12} + 2(A_{21} - A_{12})x_1]$; $ln\gamma_2 = x_1^2[A_{21} + 2(A_{12} - A_{21})x_1]$;

Raoult's law (binary mixture):

$y_iP = x_i\gamma_iP_i^{sat}$, $P = x_1\gamma_1P_1^{sat} + x_2\gamma_2P_2^{sat}$

Antoine equations (in mmHg):

$log_{10}(P^{sat}) = A - \frac{B}{T(°C)+C}$,

Model building for MLE:

The objective is to minimize the Likelihood Function such that:

$$\min_{\theta} -logL(\theta) = -\sum_{\omega\epsilon S} logf(x_\omega|\theta)$$

Where $logL(\theta)$ is the likelihood function.

One – Margules parameter: γ_1 is rearrange as: $\gamma_1 = \exp(A_{12}x_2^2)$. Here, $A_{12} = \theta$ as this is the variable we want to estimate. Transform this into $-L(\theta)$ such that:

$$-L(\theta) = -\exp\left(A_{12}\sum_{\omega=1}^{S} x_2^2\right) = -\exp\left(\theta\sum_{\omega=1}^{S} x_2^2\right)$$

Follow by a log transformation $-logL(\theta) = -\theta log(\sum_{\omega=1}^{S} x_2^2)$ to represent the likelihood function above.

Next, constraint for optimization the function in fmincon is defined in the form: $Aeq * theta = Beq$. Raoult's law and activity model is combined to satisfy this constraint such that:

$$\theta x_2^2 = ln\left(\frac{y_1 P}{x_1 P_1^{sat}}\right)$$

Two – Margules parameter: γ_1 is rearrange as:

$\gamma_1 = \exp(x_2^2[A_{12} + 2(A_{21} - A_{12})x_1])$. Here, $A_{12} = \theta_1$ and $A_{21} = \theta_2$ as these are objective variable for estimation. Using the same procedure as one-parameter model building and rearranging:

$$logL(\theta) = -\,[\theta_1\left(\sum_{\omega=1}^{S} x_2^2 - 2x_2^2\,x_1\right) + \theta_2\left(\sum_{\omega=1}^{S} 2x_2^2\,x_1\right)]$$

Similarly, constraint for optimization is:

$$[\theta_1(x_2^2 - 2x_2^2x_1) + \theta_2(2x_2^2x_1)] = ln\left(\frac{y_1 P}{x_1 P_1^{sat}}\right)$$

Model building for MLM: For Multivariate model, we want to estimate θ as follow:

$$\theta = inv(X'X)X'Y$$

To define X and Y, we want to manipulate data into the follow linear model of:

$$Y = \theta X\ for\ one - parameter\ model. Here, A_{12} = \theta$$

$$Y = \theta_1 X_1 + \theta_2 X_2\ for\ two - parameter\ model.\ \ Here, \qquad X = [X_1, X_2], A_{12} = \theta_1\ and\ A_{21} = \theta_2$$

one parameter model:

$X = x_2^2$ $\qquad$ $Y = \ln(\gamma_1)$, where $\gamma_1 = \frac{y_1 P}{x_1 P_1^{sat}}$

This can be substitute back into the MLM model to estimate θ.

For two parameter model:

$X_1 = x_2^2 - 2x_2^2x_1 \qquad X_2 = 2x_2^2x_1$,

$X = [x_2^2 - 2x_2^2x_1, 2x_2^2x_1]$ $\qquad$ $Y = \ln(\gamma_1)$ where $\gamma_1 = \frac{y_1 P}{x_1 P_1^{sat}}$

Discussion:

Non-randomized data:

The result of model prediction versus non-error generated data using two statistical models onto two activity model for two mixtures are visualized below. Overall, it seems that the approach of using linear statistical models are only accurate for linear activity model. This is reflected through 1st mixture results. For data that follows a non-linear trend, the statistical model retains the trend but fail to estimate accurately. This is reflected through 2nd mixture result.

MLE model MLM model

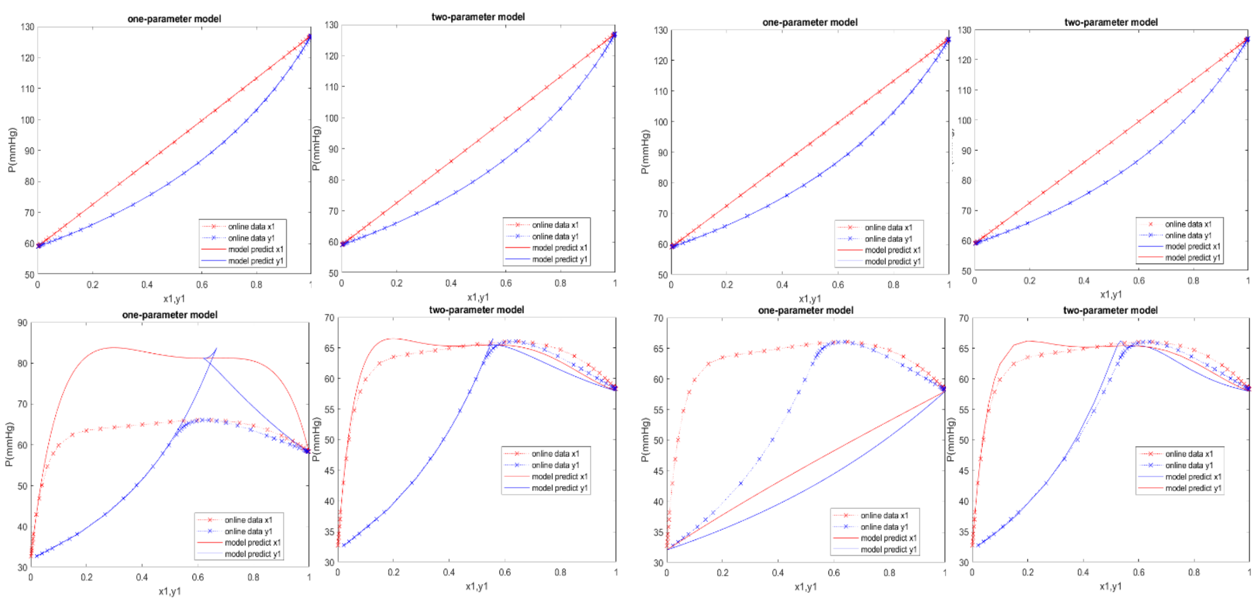

Fig 1. Methanol(1)-ethanol(2) (all top) and isopropanol(1)-water(2) (all bottom)

It should be noticed that the difference between two model's statistical approximation can be seen from their failure to approximate the 2nd mixture using one-parameter model. MLE model took on optimizing an estimate to minimize the difference while still retaining the curve trends. MLM however transform the function into a linear model, and tried to optimize as such, resulting in a linear trend similar to 1st mixture trend. In some sense, MLE would be more preferable for than MLM for modelling in this case. Despite the lack of accuracy, a general trend can still be seen with MLE.

Randomized error data:

Now, randomized error must be taken into account to reflect real-world data collection. Ebulliometer typical error provided by BulTeh User's guide estimate a $\pm 0.2\%$ standard deviation in measuring the fractional content and $\pm 0.1\%$ temperature error. To simplified error process, the temperature error will be negligible for the case, the x and y data are adjusted with 0.0002 error. The new data are re-ran multiple times with the models to check for consistency/abnormality modelling.

While accuracy, not precision, is retained for one-Margules activity model, there are great variation for two-Margules activity model when modelling both mixture. To portray the range of variation visually, the two-Margules model are ran with 100 repetition of randomization. These are shown respectively in figure 2 for MLE models and figure 3 for MLM models. Visually, MLE model seems to vary less than MLM when facing data uncertainty.

On the other hand, statistical parameter such as uniqueness, co-variance, and fisher information would reflect that MLM exhibit more significant information. As these information can only be calculated from one point, the average value of A12 (theta 1) and A21 (theta 2) from the 100 sample repetition are used to evaluate these information and are shown in table below. Both uniqueness eigenvalues have a value close to zero, which means the estimate has high variability. Interestingly, co-variance of MLE is higher than MLM, and the fisher Information eigenvalues for MLE is lower than MLM. Both of these reflect MLM estimates are sharper than MLE. Overall, for 2nd mixture, MLM model, while varies more, result in more points being closer to the actual data. MLE, while consistent and varies less visually, have less points closer to the actual data compared to MLM. Taking into account the model performance on 1st mixture, MLE overall seems like a much better statisical model to estimate activity coefficient than MLM model when facing uncertainty for its precision.

	MLE	MLM
uniqueness	[12.2476;0.7280]	[12.2248;0.7246]
Co- variance	[0.0012, -0.0013; -0.0013, 0.0184]	[3.77e-04, -4.21e-04; -4.21e-04, 0.0058]
fisher information	[8.28e+05;9.27e+03]	[8.44e+06; 2.96e+04]

Table 1: statistical information of MLE and MLM models

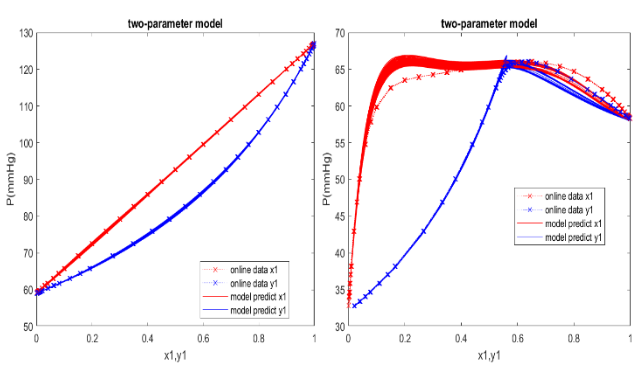

Fig 2: MLE model. 1st mixture(left), 2nd mixture(right).

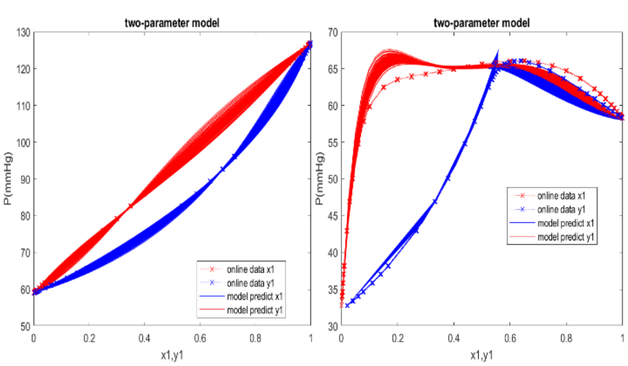

Fig 3: MLM model. 1st mixture(left), 2nd mixture(right).

Reference

http://apmonitor.com/me575/index.php/Main/VLEWilson

http://www.vle-calc.com/about_en.html

Carl T. Lira, J. Richard Elliott, Introductory Chemical Engineering Thermodynamics (New York, Prentice Hall, 2012), pp 443-470