Song, Y. H., D. W. Blair, V. J. Suminski, and W. Bartok, "Conversion of Fixed Nitrogen to N_2 in Rich Combustion," *18th Symposium (International) on Combustion*, Pittsburgh, PA, 53 (1980).
Syred, N., and J. M. Beer, "Combustion in Swirling Flows: A Review," *Combust. and Flame*, **23**, 143 (1974).
Syred, N., N. A. Chigier, and J. M. Beer, "Flame Stabilization in Recirculation Zones of Jets with Swirl," *13th Symposium (International) on Combustion*, Combustion Institute, Pittsburgh, PA, 617 (1971).
Thurgood, J. R., L. D. Smoot, and P. O. Hedman, "Rate Measurements in a Laboratory-Scale Pulverized Coal Combustor," *Comb. Sci. and Tech.*, **21**, 213 (1980).
Thurgood, J. R., "Mixing and Reaction Rate Resolution in Pulverized Coal Combustion," Ph.D. Dissertation, Brigham Young University, Provo, UT (1979).
Zeldovich, Y. A., "The Oxidation of Nitrogen in Combustion Explosions," *Acta Physiocochima*, **21**, 577 (1946).

Manuscript received January 26, 1981; revision received August 31, and accepted September 16, 1981.

Statistical Estimation of Parameters in Vapor-Liquid Equilibrium

A new procedure for the estimation of parameters in two-component vapor-liquid equilibrium is presented. It allows for measurement errors in all variables and gives better fits than most other procedures because of the use of two relations between the variables rather than one as is commonly used. An approximate and an exact solution are described and the importance of using all thermodynamic and statistical information is illustrated for the Wilson and UNIQUAC models. New alternatives for analysis of residuals are also discussed.

HUGO PATINO-LEAL
and
P. M. REILLY
Department of Chemical Engineering
University of Waterloo
Waterloo, Ontario, Canada

SCOPE

The importance of phase equilibrium prediction in chemical process design has promoted the development of various semi-mechanistic models. An important aspect in the successful application of these models is the optimal estimation of their unknown parameters from limited experimental information. In this process of estimating parameters, it is important to use all the information that is available.

One source of information is the thermodynamics of the system. This supplies the equations relating the several variables. In the past, not all equations have been used; this work illustrates the improvement that can be achieved by using simultaneously all the equations involved.

Another problem with most previous attempts has been the use of estimation criteria which are basically arbitrary. Some of these assume some of the variables to be perfectly known. Our solution is a statistical one, based on the error-in-variables model (EVM), which gives optimal parameter estimation and allows for the presence of experimental error in all measured variables, giving general and flexible solutions, easy to apply.

Once the parameter estimates have been found, it is important to have an idea of how much uncertainty is involved in them, and how well the model represents the system. We present new alternatives for these aspects.

While the statistical principle of allowing for error in all measured variables is applicable to any formulation of the problem of describing vapor-liquid equilibria, we will illustrate its use as applied to the Wilson and UNIQUAC models. We will use the illustration further to show how our approach can be used to obtain residuals to assess the applicability of the mathematical model.

As a further objective of this paper, we intend to demonstrate the importance of using all the thermodynamic information which is available to supplement that provided by the experimental results.

CONCLUSIONS AND SIGNIFICANCE

Basic thermodynamics indicates that for a binary system to have two phases in equilibrium two equilibrium conditions must be satisfied simultaneously. We have found that these two relations provide important information, all of which should be used in order to estimate efficiently the parameters of the model.

The error-in-variables procedure presented here allows for the presence of experimental random error in all the variables involved. It also allows for more than one equilibrium condition to be satisfied and does not impose any restrictions on the form of the error covariance matrix. The exact solution presented is relatively simple to program, converges quickly and does not require the storage or inversion of large matrices. An alternative (approximate) solution is also presented which has proven to give very valuable results and involves a minimum of computation. The performance of these statistically sound techniques is remarkably better than methods based on arbitrary estimation criteria.

There are several ways in which one can check how well (or badly) a particular model is able to represent the data. We have been able to provide residuals to accomplish this check by an extension of standard statistical procedures.

0001-1541-82-5770-0580-$2.00. © The American Institute of Chemical Engineers 1982.

INTRODUCTION

The importance of parameter estimation in vapor-liquid equilibrium is evident in the large number of methods proposed in the literature (Ma et al., 1969; Brinkman et al., 1974; Deak, 1974; Marina and Tassios, 1973; Holmes and van Winkle, 1970; Prausnitz et al., 1967; Bruin, 1970; Barker, 1953; Van Ness et al., 1973; Verhoeye, 1970). This importance is also reflected in the large data bases which have been published, such as the Dortmund Data Bank (DDB) by Gmehling and Onken (1975), and others like that of Hirata, et al. (1975). Almost every paper proposes a different estimation criterion involving the minimization of the sums of squares of some deviations between adjusted and measured values of the pressures, vapor mole fraction of one or several components, activity coefficients, etc. Also there appears to exist confusion even among the latest works as to how many variables and relations there are or should be used (Sutton and MacGregor, 1977; Kemeny and Manczinger, 1978; Anderson et al., 1978).

PROBLEM OF PARAMETER ESTIMATION

For a given binary system, say Propanol-Water, a typical set of experimental data consists of measurements of the pressure, temperature, and mole fraction of one of the components (say, component 1) for the vapor phase and for the liquid phase for each of the $i = 1, \ldots n$ experiments. Each of these measurements contains error more or less; consequently, we distinguish between the measured and the true values of the variables as:

$$\begin{aligned} P_i &= \xi_{1i} + \epsilon_{1i} \\ y_{1i} &= \xi_{2i} + \epsilon_{2i} \\ i &= 1, \ldots n \\ x_{1i} &= \xi_{3i} + \epsilon_{3i} \\ T_i &= \xi_{4i} + \epsilon_{4i} \end{aligned} \tag{1}$$

There are two more variables which usually are not measured. They are the mole fraction of the second component in the vapor phase and in the liquid phase. From material balances, it follows that

$$\begin{aligned} \xi_{5i} &= 1 - \xi_{2i} \\ \xi_{6i} &= 1 - \xi_{3i} \end{aligned} \tag{2}$$

From thermodynamics, for a binary system to be in equilibrium the following conditions must apply simultaneously

$$\begin{aligned} \phi_{1i}\, \xi_{2i}\, \xi_{1i} &= \gamma_{1i}\, \xi_{3i}\, f^{\circ}_{1i} \\ \phi_{2i}\, \xi_{5i}\, \xi_{1i} &= \gamma_{2i}\, \xi_{6i}\, f^{\circ}_{2i} \end{aligned} \tag{3}$$

Numerical values of the fugacity coefficients and standard state fugacities can be obtained as functions of the variables which are measured. The activity coefficients are given by expressions like the Wilson or UNIQUAC equations which each contain two unknown parameters θ_1, θ_2, whose values are to be estimated. We may represent Eqs. 3 in the form

$$\begin{aligned} g_{1i}(\xi_{1i}, \xi_{2i}, \xi_{3i}, \xi_{4i}, \theta_1, \theta_2) &= 0 \\ g_{2i}(\xi_{1i}, \xi_{2i}, \xi_{3i}, \xi_{4i}, \theta_1, \theta_2) &= 0 \end{aligned} \tag{4}$$

These relations are nonlinear in both the ξ's and θ's.

For n sets of measurements, represented by the left sides of Eqs. 1, the information from which the parameter estimates must be derived is given by Eqs. 1 and 4. Our approach for parameter estimation is a statistical one which may be considered either a likelihood method or a Bayesian one as suits the reader's taste. That is, our point parameter estimates are optimal in the sense that they may alternatively be considered maximum likelihood estimates or modes of the posterior parameter distributions. In terms of sums of squared differences this means that they are obtained by minimizing the weighted sum of squared differences between the observed and fitted values of all measured quantities. The weighting is according to the precision of the measurements. This sum of squares criterion was not chosen for convenience. It is the result of a search for a maximum likelihood or Bayesian estimate for the parameters.

Accordingly, as equations 1 indicate, we take into account the error in the measurement of all variables. For this reason, our model is called by statisticians an error in variables model (EVM). The full description of our solution of the EVM problem is found in Reilly and Patino-Leal (1981).

In using this approach, we avoid the arbitrariness of making a basic choice of estimation criterion and at the same time overcome the limitation of the conventional least squares approach which disregards error in all variables but one. In effect, our procedure is equivalent to a generalization of the approaches of Sutton and MacGregor (1977), Anderson et al. (1978), and others. This generalization allows for a more flexible error structure than the previous papers. More important, it allows for any number of equilibrium conditions to apply simultaneously, so that the procedure can be extended to multicomponent equilibrium.

Sometimes "incomplete" sets of data, such as isothermal X-P data, or isobaric X-T data are reported. The EVM procedure can also be applied to these sets of data.

NUMBER OF RELATIONS

In any situation where parameters are being estimated statistically, it is obviously important to make measurements on as many relevant variables as possible. It is also important, as Box and Draper (1965) show, to use as many relations as possible which involve the parameters to be estimated. In equilibrium systems, there is a connection between the number of variables measured and the required number of relations between them which may be derived from the Phase Rule.

The Phase Rule states that the degrees of freedom of an equilibrium system is given by $m + 2 - r$. If the system is completely modelled mathematically and we want to use the equations of the model to determine the values of a total of v variables, including measured or specified ones, we must have the following number of equations in the model:

$$E = v - (m + 2 - r) \tag{5}$$

This follows because variables in number equal to the degrees of freedom may be specified arbitrarily; for each one beyond these, an equation is needed to describe it as a function of the free variables.

If, on the other hand, the values of v variables have been measured, possibly several times each, and we are attempting to obtain parameter estimates, efficient estimation requires that we use E-independent equations relating the variables, as specified by Eq. 5, because in estimating parameters we in effect use our model to calculate fitted values of the variables and then compare them with the measured ones.

If we use fewer than E equations, it is always possible by a least squares method to obtain unique parameter estimates if only the number of sets of observed variables exceeds the number of parameters. However, inefficient estimates are obtained, because their values are not constrained by all the applicable relations. We will demonstrate this later.

If more than E equations are used, the excess ones are redundant. It is not possible to obtain useful parameter estimates under this condition.

In our case, Eq. 5 specifies two relations. In fact, we have exactly two (Eq. 4); consequently, we can be assured of efficient estimation. Many writers, such as Sutton and MacGregor (1977) and Kemeny and Manczinger (1978) have used for one reason or another fewer equations than specified by Eq. 5 and consequently have obtained less than efficient estimation.

Box, Hunter, MacGregor, and Erjavec (1973) demonstrate the need, when using the method of Box and Draper for parameter estimation, to classify equations as those which represent linear

relations among the data as against those which represent linear relations among their expectations. They point out that those of the former type lead to serious instability in parameter estimation and consequently must be eliminated.

In our situation, no such restriction is required because we assume the covariance matrix of the errors to be known. We believe that in practice it is little hardship to perform a special experiment or to design into the main experiment replicate measurements on each variable from which an estimate of the covariance matrix may be obtained. In some cases the covariance matrix is known approximately from previous experience. Because of this our solution is valid even when there are linear relations among the data, corresponding to a known but singular error covariance matrix. The Box-Draper method applies where the error covariance matrix is unknown and it is this fact which makes it sensitive to singularity of the matrix.

USE OF ERROR-IN-VARIABLES MODEL

We have n sets of measurements represented by the quantities on the left of Eq. 1 along with Eqs. 1 and 4 from which to estimate the parameters θ_1 and θ_2. The problem is seriously complicated by the presence of the nuisance parameters $\xi ji;\ j = 1, 2, \ldots 4;\ i = 1, 2, \ldots n$. An estimation scheme for θ_1, θ_2 and the other $4n$ parameters has been proposed previously by Anderson et al. (1978) by extending an approach similar to that of Britt and Luecke (1973), assuming measurement error to be independent. Sutton and MacGregor (1977), and Kemeny and Manczinger (1978) have proposed the use of the error propagation method (Box, 1970). Sutton and Mac Gregor used only one equilibrium condition, whereas Kemeny and Manczinger used one or two depending on whether three or four variables were used.

If we consider the errors ϵ_i to be normally and independently distributed for each i, with mean vector zero and a covariance matrix V, then the likelihood of the observations z is given by

$$Df(\underline{Z}|\underline{\xi}_i,\underline{\theta}) \alpha \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\underline{z}_i - \underline{\xi}_i)' V^{-1}(\underline{z}_i - \underline{\xi}_i)\right\} \quad (6)$$

and this is subject to the functional relations (equilibrium conditions) given in Eq. 4 which may be linearized as follows by a Taylor's series expansion at each $\hat{\underline{\xi}}_i$.

$$g(\hat{\underline{\xi}}_i,\underline{\theta}) + \underline{B}_i(\underline{\xi}_i - \hat{\underline{\xi}}_i) = 0 \quad (7)$$

The vectors $\hat{\underline{\xi}}_i$ $(i = 1, \ldots, n)$ are the points in the space of the ξ's at which we make the linearizations. The possibilities of what these may be will be discussed shortly.

We now eliminate the ξ_i by integrating them out rather than estimating them. Assume a non-informative parameter distribution on ξ_i from Jeffreys' rule as applied by Box and Tiao (1973), and a completely uniform prior on $\underline{\theta}$. It can then be shown that, after integrating out the vectors ξ_i, the marginal posterior distribution for $\underline{\theta}$ is given by

$$Df(\underline{\theta}|\underline{Z}) \alpha \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} Q_i\right\} \quad (8)$$

where

$$Q_i = [g(\hat{\underline{\xi}}_i,\underline{\theta}) + \underline{B}_i(\underline{z}_i - \hat{\underline{\xi}}_i)]'(\underline{B}_i\underline{V}\underline{B}_i')^{-1}[g(\hat{\underline{\xi}}_i,\underline{\theta}) + \underline{B}_i(\underline{z}_i - \hat{\underline{\xi}}_i)]$$

For details see Patino-Leal (1979), Reilly and Patino-Leal (1981). This expression does not depend on the true values of the measured variables ξ_i. It is a function only of $\underline{\theta}$ and the linearization points $\hat{\underline{\xi}}_i$. Alternatives for the $\hat{\underline{\xi}}_i$ will now be discussed.

APPROXIMATE AND EXACT SOLUTIONS

Linearizing at the Measured Values z_i (Approximate Solution)

If we make $\hat{\underline{\xi}}_i = z_i$ for all $i = 1, \ldots n$, the posterior distribution given in Eq. 8 becomes

$$Df(\underline{\theta}|\underline{Z}) \alpha \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}[g(\underline{z}_i,\underline{\theta})]'(\underline{B}_i\underline{V}\underline{B}_i')^{-1}[g(\underline{z}_i,\underline{\theta})]\right\} \quad (9)$$

Although this is an approximate solution, it is very convenient and easy to apply. To get the point estimates, we have to do only a two parameter search by minimizing the summation inside the exponential. We will describe some simulated experiments in which this solution provided very valuable results. This alternative corresponds to a generalization of the error propagation method for two simultaneous relations.

Exact Solution

An improvement over the previous solution can be made since the value of the posterior probability given in Eq. 9 is not the actual one, but only an approximation. We can find the values $\hat{\underline{\xi}}_i$ where one should linearize the Eqs. 4 so that the solution given in Eq. 8 is exact. It can be shown (Reilly and Patino-Leal, 1981) that an exact linearization for Eqs. 4 may be found by applying iteratively the following equation for each $i = 1, 2, \ldots n$.

$$\hat{\underline{\xi}}_i^{(j+1)} = \underline{z}_i - \underline{V}\underline{B}_i'(\underline{B}_i\underline{V}\underline{B}_i')^{-1}[\underline{g}(\hat{\underline{\xi}}_i^{(j)},\theta) + \underline{B}_i(\underline{z}_i - \hat{\underline{\xi}}_i^{(j)})] \quad (10)$$

$$j = 1, 2, \ldots$$

To start the iteration $\hat{\underline{\xi}}_i^{(1)} = z_i$ is usually effective and the calculation should be continued until Q_i as defined in Eq. 8 is stable to the required degree. We have applied this procedure and found that it coverges rather quickly, after usually requiring no more than five iterations at each i. Notice that we are only handling small matrices, and we only have to invert a 2×2 matrix. Also no restrictions are placed on the form of $\underline{V}$, not even that it be nonsingular. Anderson, Abrams and Grens (1978), for example, require $\underline{V}$ to be diagonal.

We now present some simulated experiments in which the approximate solution given in Eq. 9 is compared with several least squares procedures. After that, a comparison will be made between the approximate and exact EVM solutions. Full details for calculations are given in Appendix 1.

SIMULATION OF EXPERIMENTS

To assess the importance of using all information available when estimating the parameters involved in a VLE model, simulated experiments were carried out for both the Wilson and the UNIQUAC models.

Wilson Equation

A simulation of experimental data was made for the isothermal (1) Propanol-(2) Water System. The true values of the Wilson parameters were set to be $\theta_1 = 3{,}768$ J/mol and $\theta_2 = 5{,}024$ J/mol, and the number of observations on each variable was 10. Fixing the temperature and liquid composition, the vapor composition and the pressure were obtained by solving a system of two nonlinear equations with two unknowns. Table 1 presents this set of ten "perfect" observations. We then introduced errors normally and independently distributed, with mean zero and the following standard deviations:

- 0.005 mol fraction for the liquid composition
- 0.015 mol fraction for the vapor composition
- 0.001 atm. for the pressure
- 0.1 K for the temperature

Twenty such sets of ten observations were generated, and from each set the parameter estimates were found using different methods:

TABLE 1. SET OF 10 "PERFECT" OBSERVATIONS WHERE THE VALUES OF THE WILSON PARAMETERS ARE θ_1 = 3,768 J/mol, θ_2 = 5,024 J/mol

Pressure (kPa)	x_1 (mol fraction)	y_1 (mol fraction)	Temp. (K)
81.85	0.05	0.4123	354.15
91.11	0.15	0.4844	354.15
94.72	0.25	0.5159	354.15
97.46	0.35	0.5461	354.15
99.71	0.45	0.5801	354.15
101.43	0.55	0.6206	354.15
102.42	0.65	0.6702	354.15
102.40	0.75	0.7330	354.15
100.96	0.85	0.8151	354.15
97.61	0.95	0.9271	354.15

TABLE 2. ESTIMATES OF WILSON EQUATION PARAMETERS FROM 20 SIMULATED EXPERIMENTS

Method	Average θ_1	Average θ_2	RMSD θ_1	RMSD θ_2
I	3,643.31	5,052.8	960.68	283.48
II	3,611.53	5,023.24	615.86	93.74
III	3,907.91	5,096.43	1,178.43	284.02
IV	3,761.04	5,025.55	211.51	25.08
True Values	3,768	5,024		

I. Method of Gmehling and Onken (1977)
II. EVM using only the equilibrium condition for propanol
III. EVM using only the equilibrium condition for water
IV. EVM using both equilibrium conditions

Method I uses one of the many minimization criteria which have previously been proposed and was used in the construction of a very large experimental data base. (Dortmund Data Base, Gmehling and Onken, 1975). It minimizes the relative differences between the observed and adjusted activity coefficients, for both components. Table 2 summarizes the results, where RMSD refers to the root mean squared deviation for the 20 simulations.

One can notice a large difference between equilibrium conditions on comparing II and III, but more important, the parameter estimation is substantially improved when both are used simultaneously (IV). This is an illustration of the inefficient estimation previously referred to when fewer equations than those prescribed by Eq. 5 are used. Notice also the very large improvement when using IV over I.

UNIQUAC Equation

A similar simulation was conducted for the UNIQUAC model. Here an isothermal (1) Methanol-(2) Toluene system was chosen, with the true values of the UNIQUAC parameters being θ_1 = −418.7 J/mol, θ_2 =4,186.7 J/mol. The set of 13 "perfect" observation shown in Table 3 was found in the same way as those of Table 1. Forty sets of 13 observations were generated using the same error structure described previously. Four methods, similar to those of the last section were compared and the results of the 40 simulated experiments are summarized in Table 4 and illustrated in Figures 1A to 1D.

Again, a very substantial improvement can be observed when we use EVM with the two equilibrium conditions (IV), over the use of the least squares procedure. The difference between the individual equilibrium conditions, and the advantage of using both simultaneously can also be observed (II, III, IV). As one can see from figures 1(a) to 1(d), an important increase in precision of the point

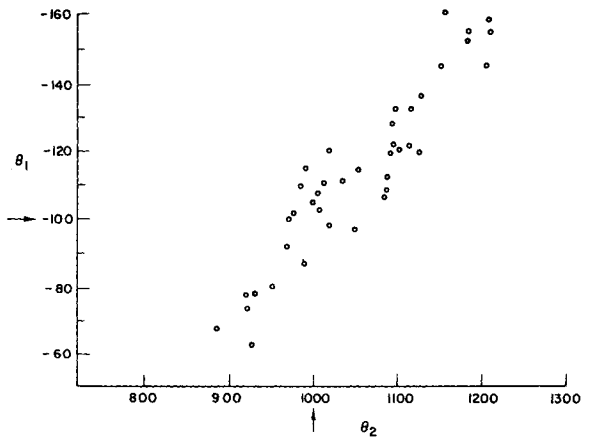

Figure 1(a). Simulation results: Method I.

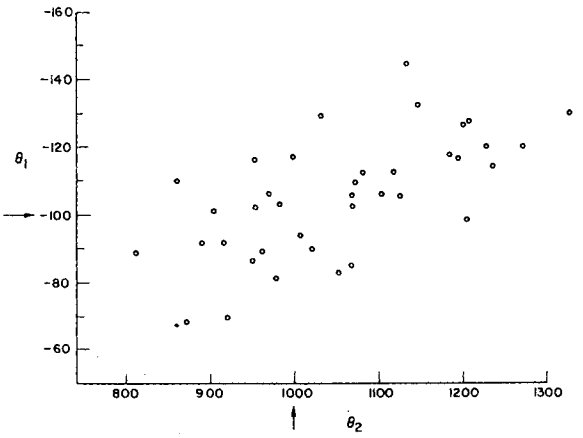

Figure 1(b). Simulation results: Method II.

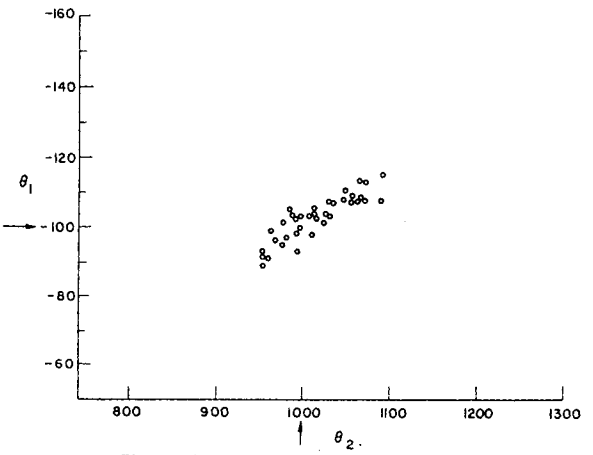

Figure 1(c). Simulation results: Method III.

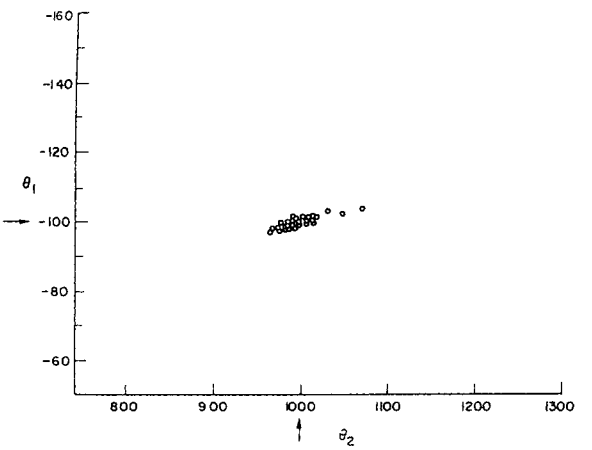

Figure 1(d). Simulation results: Method IV.

TABLE 3. SET OF 13 "PERFECT" OBSERVATIONS WHERE THE VALUES OF THE UNIQUAC PARAMETERS ARE $\theta_1 = -418.7$ J/mol, $\theta_2 = 4{,}186.7$ J/mol

Pressure (kPa)	Temp. (K)	ξ_3 (mol fr.)	ξ_2 (mol fr.)
57.87	338.15	0.05	0.6187
73.19	338.15	0.10	0.7021
81.21	338.15	0.15	0.7348
86.01	338.15	0.20	0.7525
91.55	338.15	0.30	0.7726
94.91	338.15	0.40	0.7859
97.48	338.15	0.50	0.7979
99.75	338.15	0.60	0.8113
101.38	338.15	0.70	0.8289
103.78	338.15	0.80	0.8553
104.48	338.15	0.85	0.8747
104.81	338.15	0.90	0.9014
104.45	338.15	0.95	0.9399

TABLE 4. ESTIMATES OF UNIQUAC PARAMETERS FROM 20 SIMULATED EXPERIMENTS

Method	Average θ_1	Average θ_2	RMSD θ_1	RMSD θ_2
I	−476.28	4,415.96	119.94	428.54
II	−437.68	4,436.81	83.90	624.61
III	−429.68	4,240.33	28.09	169.52
IV	−420.22	4,179.16	6.80	85.63
	−418.7	4,186.7		

estimates can be achieved by using EVM with both equilibrium conditions.

Exact vs. Approximate Solution

We have compared the use of the approximate vs. the exact solution for some of these simulated sets of data. The point estimates are usually very close (around one or two percent difference, although sometimes the difference may be larger). The use of the exact solution offers a marginal improvement over the approximate solution. When comparing the difference: (a) between parameter estimates of these two EVM solutions; with (b) the difference between the use of the least squares procedure (Method I) and the approximate EVM solution, we found the difference (a) to be usually less than about one-tenth of (b). However, we found one instance in which (a) was one-fourth of (b). We recommend that the exact procedure be ordinarily used because it is easy to program and as mentioned previously it converges rapidly. If, however, this extra step is not desired, the approximate solution gives a fairly good answer, especially when compared to ordinary least squares procedures.

ANALYSIS OF RESIDUALS

It is of great importance to analyse how well a particular set of experimental data fits the equation which is used (UNIQUAC, Wilson, or others). This analysis shows if there is any trend or systematic deviation which might indicate a lack of fit of the equation, or the presence of nonconstant error variance for example. The usual consistency tests used in VLE can only indicate a global or general quality of the data, and do not refer directly to the fit of the data *to the particular model* being considered (e.g., Wilson, UNIQUAC, NRTL, etc.).

It can be shown (Reilly and Patino-Leal, 1981) that several sets of meaningful residuals can be obtained in this case. They can be used to illustrate the general adequacy of the model. The most meaningful of these are the two sets which show the departure of each measured vector of variables from the two Eqs. 4. The residuals for the jth equilibrium condition are given by

$$\frac{[g(z_i,\hat{\theta}) + \underline{B}_i(z_i - \hat{\xi}_i)]_j}{[\underline{B}_i \underline{V} \underline{B}_i']_{jj}} \quad j = 1, 2 \quad i = 1, 2, \ldots n \qquad (11)$$

These residuals represent a form of directed distances in the space of the ξ's from the observations to each one of the surfaces representing the equilibrium conditions. They contain as a special case the usual least squares residuals which are obtained by using the appropiate $\underline{V}$. They do not necessarily require the use of the estimated true values of the observations, and they express the overall quality of the fit of the data to the model. This is a different approach from those reported previously (e.g., Anderson, et al., 1978). There, the true values of the observations are estimated, and then residuals are found for one or several of the variables by subtracting the estimated true values from the measured ones.

There is a further set of residuals which represents the joint lack of fit of the two equilibrium conditions to the data. This consists of the quantities $Q_i^{1/2}$, $i = 1, 2, \ldots$ which represent the absolute values of the distances in the space of the ξ's from the measured points to the intersection of the surfaces representing the two equilibrium conditions. It is more difficult to associate a sign with these quantities than with the residuals in Eq. 10 because their orientation in ξ-space is more complicated.

UNCERTAINTY IN PARAMETER ESTIMATES

The point estimates obtained for the two parameters of the model must be supplemented with information about the amount of uncertainty associated with these parameters. This can be found from the joint posterior contours (equivalent to confidence regions). To get these contours we have several alternatives according to whether if we want an exact or an approximate contour. Reilly (1976) presents a numerical computation procedure to get the exact contours; and Box and Cox (1964) present a chi-squared approximation which gives the correct shape of the posterior distribution at an approximate probability level. Another approximation is often useful when the true posterior contours are known to be approximately elliptical as is usually true in our case. This involves an approximation of the posterior probability density function by a multivariate normal distribution (Pritchard et al., 1977). One only needs to find the points θ_1, θ_2 on the contour that satisfy the equation

$$h_{11}(\theta_1 - \hat{\theta}_1) + 2h_{12}(\theta_1 - \hat{\theta}_1)(\theta_2 - \hat{\theta}_2) + h_{22}(\theta_2 - \hat{\theta}_2)^2 = \chi^2_{2,0.95}$$

The quantities h_{11}, h_{12} and h_{22} are easy to find and they are sometimes a by-product of the optimization procedure used, like any of the variable metric methods presented by Bard (1974).

ANALYSIS OF AN ACTUAL SET OF DATA

Consider the data presented by Wilson and Simons (1952) on the system (1)2-Propanol, (2) Water. (Gmehling and Onken, 1975, p. 334). This set of data passed both thermodynamic consistency tests and Gmehling and Onken report 2,115.08 J/mol and 5,492.78 J/mol as point estimates for the Wilson parameter. Using their method we obtained a slightly different answer: 2,015.5 J/mol and 5,660.4 J/mol. The difference between these two sets is due mainly to the assumption of Gmehling and Onken that the fugacity coefficients are equal to one. It is important to try to avoid unnecessary approximations like this one, since the effect on point estimates might be larger than expected. See Appendix 1 for details of the calculation procedure.

The use of the approximate EVM solution given by Eq. 9 gave drastically different estimates: 5,043.7 J/mol and 5,331.3 J/mol. EVM point estimates using the exact solution (Eqs. 8 and 10) were 6,131.8 J/mol and 5,272.7 J/mol. In this particular case there is a substantial difference between the approximate and exact solutions.

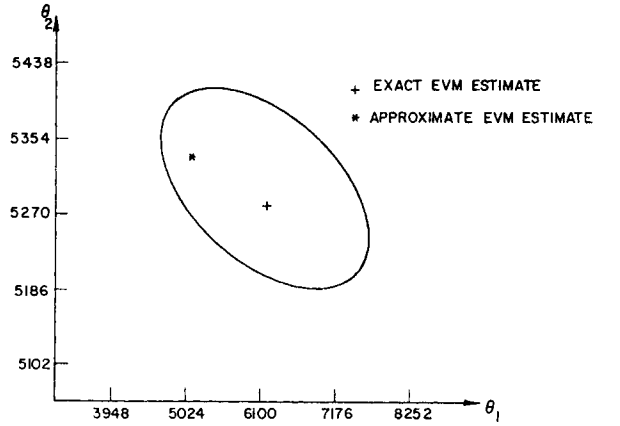

Figure 2. Approximate 95% posterior probability contour for Wilson parameters: data of Wilson and Simons.

Figure 2 presents an approximate 95% posterior probability contour for the exact solution. The EVM point estimates using the approximate solution (Eq. 9) is inside the ellipse; however, the estimate using the method of Gmehling and Onken is outside.

To illustrate with real data the consistency of the EVM solution compared to the solution of Gmehling and Onken, we prepared parameter estimates based on ten selected subsets of ten observations each from the complete data set of 26 observations. Each of these subsets covered most of the range of the complete set. The point estimates are summarized in Table 5. The smaller RMSD's for EVM show a more consistent estimation. This behavior was also evident in the simulations we presented previously.

As mentioned previously, several arbitrary estimation criteria have been proposed in the past. We compare here the results obtained using some of these criteria with our results. Consider first the method of Gmehling and Onken. Let

$$SS(P) = \sum_{i=1}^{n} (P_i - \hat{\xi}_{1i})^2$$

$$DD(P) = SS(P)/\sigma_P^2$$

and similarly for y_1, x_1, T. The SS's are simply the sum of squares of the differences between observed and adjusted values for the variables, while the DD's (directed distances) take into consideration how much error is introduced in each variable. Table 6 summarizes the results. The Gmehling and Onken method produces a smaller $SS(y_1)$ than EVM, but the rest are larger. More important, the sum of DD's for all four variables is three times larger for Gmehling and Onken's estimate than for the EVM estimate.

TABLE 5. ESTIMATES OF WILSON PARAMETERS FOR SUBSETS OF 10 OBSERVATIONS. EXPERIMENTAL DATA BY WILSON AND SIMONS

	Gmehling and Onken	E.V.M.
Average θ_1	2,333.9	5,500.1
Average θ_2	5,673.1	5,302.8
RMSD θ_1	971.9	807.2
RMSD θ_2	267.9	52.2

TABLE 6. COMPARING THE CONTRIBUTIONS OF EACH VARIABLE TO THE GLOBAL OBJECTIVE, IF WE USE THE POINT ESTIMATES BY GMEHLING AND ONKEN OR THE EVM POINT ESTIMATES

	E.V.M.		Gmehling and Onken	
	SS (·)	DD (·)	SS (·)	DD (·)
P	0.000216	23.7653	0.0003638	9.5712
y	0.005347	233.76	0.0021535	1,061.76
x	0.005844	216.12	0.026544	363.84
T	0.037632	3.7632	0.078504	7.85
Total	...	477.4	...	1,443.

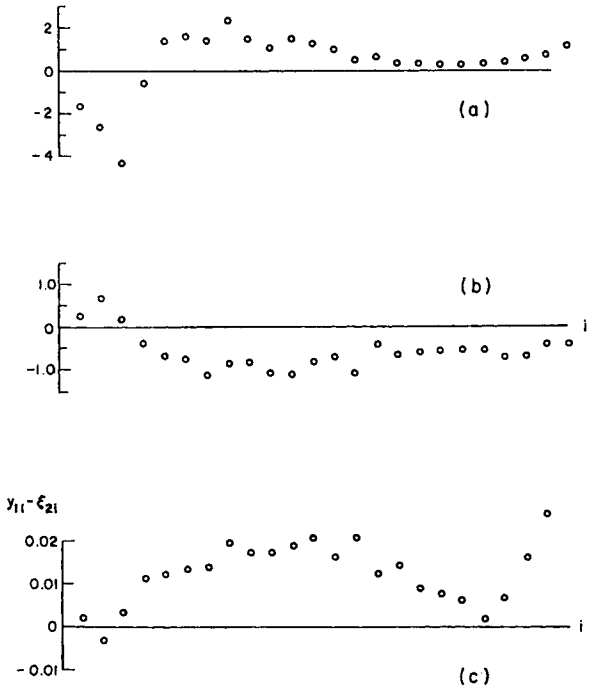

Figure 3. Analysis of residuals for experimental data. 3(a) and 3(b) are EVM residuals for the two equilibrium conditions. 3(c) gives EVM residuals for y_1 alone.

Finally, Figures 3(a), 3(b), 3(c) present the analysis of residuals. The first two figures give the EVM residuals for each equilibrium condition as proposed in Eq. 11. Figure 3c gives the EVM residuals as suggested by Anderson et al. (1978) calculated by the difference between the observed and adjusted value for y_1. They all seem to show some systematic deviation.

ACKNOWLEDGMENT

The authors acknowledge the assistance of grants from the Natural Science and Engineering Research Council of Canada. Our gratitude goes to Dr. Gary Blau and reviewers for their valuable ideas and suggestions.

NOTATION

B = (mxv) matrix of partial derivatives of each equilibrium condition with respect to each variable
E = smallest number of equations required to uniquely describe the thermodynamic system
f_i° = Standard-state fugacity for component i
g_i = equilibrium condition for component i
g = vector of equilibrium conditions $(mx1)$
h_{11}, h_{12}, h_{22} = elements of the Hessian matrix of the posterior distribution at $\underline{\theta} = \hat{\underline{\theta}}$
m = number of components in the system
n = number of experiments
P_i = measured pressure for ith experiment
Q_i = quadratic form inside the exponential for the marginal posterior distribution of $\underline{\theta}$ defined in Eq. 8
r = number of phases in the system
T_i = measured temperature for the ith experiment
v = number of variables being measured
$\underline{V}$ = error covariance matrix (vxv)
x_{1i} = measured mole fraction of component 1 in the liquid phase for the ith experiment
y_{1i} = measured mole fraction of component 1 in the vapor phase for the ith experiment
z_i = $(P_i, x_{1i}, y_{1i}, T_1)'$
$\underline{Z}$ = $(z_1, z_2, \ldots, z_n)'$

Greek Letters

γ_{ji} = activity coefficient of the jth component for the ith experiment

ϵ_{ji} = measurement error in the jth variable for the ith experiment

$\underline{\epsilon}_i = (\epsilon_{1i},\epsilon_{2i},\epsilon_{3i},\epsilon_{4i})'$

θ_1,θ_2 = parameters for the Wilson model (usually denoted in previous works by $\lambda_{11}-\lambda_{12}$, $\lambda_{12}-\lambda_{22}$; or for the UNIQUAC model (usually denoted by $u_{11}-u_{12}$, u_{12})

$\hat{\theta}_1,\hat{\theta}_2$ = EVM point estimates of θ_1, θ_2

ξ_{ji} = true value of the ith observation on the jth variable

$\hat{\xi}_{ji}$ = estimated value of the ith observation on the jth variable

$\underline{\xi}_i$ = $(\xi_{1i},\xi_{2i},\xi_{3i},\xi_{4i})'$, and similarly for $\hat{\underline{\xi}}_i$

σ_P^2 = error variance for the pressure

ϕ_{ji} = fugacity coefficient of component j for the ith experiment

$\chi^2_{0.95,2}$ = value of the chi-squared distribution for 2 degrees of freedom and 95% probability

APPENDIX 1

Details of the computations for estimating the parameters of a given Vapor-liquid Equilibrium model.

This is only an illustrative example. The thermodynamic details might vary from one example to another.

In this example we have used a variable metric method to optimize with respect to the Wilson or UNIQAC equation parameters. Any desired method could be used but as described in Reilly and Patino-Leal the variable metric method has the advantage that it provides as a by-product information about the degree of uncertainty associated with the parameter estimates.

Preliminary Information

The following information is necessary for the computations:

a) The saturation pressure of each pure component (P_K^s). This may be obtained either experimentally or by the use of correlations, such as the Antoine Equation, as described by Prausnitz et al. (1967).

b) Liquid molar volumes. These may be obtained from correlations, such as that of Yen and Woods (1966).

c) Second virial coefficients. They can be obtained following the method described by Hayden and O'Connell (1975). Also, Fredenslund, Gmehling and Rasmussen (1977) tabulate these for over seventy compounds.

d) The covariance matrix of the measurement errors is preferably obtained from replicate measurements as described by Reilly and Patino-Leal but may possibly be based on the judgment of the experimenter.

The Equations Involved

Defining Fugacity Coefficients and Standard State Fugacities. The following relations are necessary. One is the expression of the fugacity coefficient for component K at the ith experimental condition Φ_{Ki}, as a function of the second virial coefficients S_{Kj} and of the pressure, temperature and vapor phase composition:

$$ln\ \phi_{Ki} = \frac{\xi_{1i}}{R\xi_{4i}}\left(2\sum_{j=1}^{m} y_j S_{Kj} - \sum_{l=1}^{m}\sum_{j=1}^{m} y_l y_j S_{lj}\right)$$

(The values of y in this equation are actually the true values of the vapor phase composition, not the measured values. The use of y instead of ξ_2,ξ_5, etc., is for notational convenience.)

Another expression that is needed is the standard-state fugacity f°_{Ki} as a function of the saturation pressure P_K^s and the fugacity coefficient at the saturation pressure Φ_K^s:

$$f^\circ_{Ki} = \phi^S_{Ki} P^S_{Ki} \exp\left\{-\left(\frac{(\xi_{1i}-P^S_{Ki})V_{LK}}{R\xi_{4i}}\right)\right\}$$

The exponential term is referred to as the Poynting correction.

Model Equations. The two relations for a binary system (given by Eqs. 3):

$$\phi_{1i}\xi_{2i}\xi_{1i} = \gamma_{1i}\xi_{3i}f^\circ_{1i}$$
$$\phi_{2i}\xi_{5i}\xi_{1i} = \gamma_{2i}\xi_{6i}f^\circ_{2i}$$

are expressed as:

$$g_1(\xi_{1i},\xi_{2i},\xi_{3i},\xi_{4i},\theta_1,\theta_2) = \phi_{1i}\xi_{2i}\xi_{1i} - \gamma_{1i}\xi_{3i}f^\circ_{1i} = 0$$

$$g_2(\xi_{1i},\xi_{2i},\xi_{3i},\xi_{4i},\theta_1,\theta_2) = \phi_{2i}(1-\xi_{2i})\,\xi_{1i} - \gamma_{2i}(1-\xi_{3i})f^\circ_{2i} = 0$$

and these two are linearized with respect to ξ at the points $\hat{\underline{\xi}}$ as

$$\underline{g}(\hat{\underline{\xi}}_i,\underline{\theta}) + \underline{B}_i(\underline{\xi}_i - \hat{\underline{\xi}}_i) = \underline{0}$$

where

$$\underline{B}_i = \begin{bmatrix} \frac{\partial g_1}{\partial \xi_1} & \frac{\partial g_1}{\partial \xi_2} & \frac{\partial g_1}{\partial \xi_3} & \frac{\gamma g_1}{\partial \xi_4} \\ \frac{\partial g_2}{\partial \xi_1} & \frac{\partial g_2}{\partial \xi_2} & \frac{\partial g_2}{\partial \xi_3} & \frac{\partial g_2}{\partial \xi_4} \end{bmatrix}$$

and

$$\frac{\partial g_1}{\partial \xi_1} = \frac{\partial \phi_1}{\partial \xi_1}\xi_2\xi_1 + \phi_1\xi_2 - \gamma_1\xi_3\frac{\partial f^\circ_1}{\partial \xi_1}$$

$$\frac{\partial g_2}{\partial \xi_1} = \frac{\partial \phi_2}{\partial \xi_1}(1-\xi_2)\xi_1 + \phi_2(1-\xi_2) - \gamma_2(1-\xi_3)\frac{\partial f^\circ_2}{\partial \xi_1}$$

$$\frac{\partial g_1}{\partial \xi_2} = \frac{\partial \phi_1}{\partial \xi_2}\xi_2\xi_1 + \phi_1\xi_1 - \gamma_1\xi_3\frac{\partial f^\circ_1}{\partial \xi_2}$$

$$\frac{\partial g_2}{\partial \xi_2} = \frac{\partial \phi_2}{\partial \xi_2}(1-\xi_2)\,\xi_1 - \phi_2\xi_1 - \gamma_2(1-\xi_3)\frac{\partial f^\circ_2}{\partial \xi_2}$$

$$\frac{\partial g_1}{\partial \xi_3} = \frac{\partial \gamma_1}{\partial \xi_3}\xi_3 f^\circ_1 - \gamma_1 f^\circ_1$$

$$\frac{\partial g_2}{\partial \xi_3} = \frac{\partial \gamma_2}{\partial \xi_3}(1-\xi_3)f^\circ_2 + \gamma_2 f^\circ_2$$

$$\frac{\partial g_1}{\partial \xi_4} = \frac{\partial \phi_1}{\partial \xi_4}\xi_2\xi_1 - \frac{\partial \gamma_1}{\partial \xi_4}\xi_3 f^\circ_1 - \gamma_1\xi_3\frac{\partial f^\circ_1}{\partial \xi_4}$$

$$\frac{\partial g_2}{\partial \xi_4} = \frac{\partial \phi_2}{\partial \xi_4}(1-\xi_2)\,\xi_1 - \frac{\partial \gamma_1}{\partial \xi_4}(1-\xi_3)\,f^\circ_1 - \gamma_1(1-\xi_3)\frac{\partial f^\circ_2}{\partial \xi_4}$$

The expressions for γ_1 and γ_2 depend on the model we are assuming (Wilson, UNIQUAC, etc.) They are in terms of ξ_3 and ξ_4, along with the unknown parameters we want to estimate (θ_1,θ_2) and some other known parameters, such as pure-component molecular structure constants for the UNIQUAC model, or v_{LK} for the Wilson equation.

For example, for the Wilson equation, the expression for γ_1 is given by

$$\gamma_1 = \exp\left\{-ln\left[\xi_3 + (1-\xi_3)\left(\frac{v_{L2}}{v_{L1}}\exp\left(\frac{-\theta_1}{R\xi_4}\right)\right)\right] + (1-\xi_3)\right.$$
$$\left.\times\left[\frac{\frac{v_{L2}}{v_{L1}}\exp\left(-\frac{\theta_1}{R\xi_4}\right)}{\xi_3 + (1-\xi_3)\frac{v_{L2}}{v_{L1}}\exp\left(-\frac{\theta_1}{R\xi_4}\right)} - \frac{\frac{v_{L1}}{v_{L2}}\exp\left(-\frac{\theta_2}{R\xi_4}\right)}{\xi_3\frac{v_{L1}}{v_{L2}}\exp\left(-\frac{\theta_2}{R\xi_4}\right) + (1-\xi_3)}\right]\right\}$$

Algorithms

As described in the main text of this paper, there are basically two alternatives for solution: an exact and an approximate solution, depending on the criteria for linearization.

Approximate Solution.

i) Calculate the preliminary information described on part 1 of this appendix for all experimental conditions (i.e., for $i = 1, \ldots, n$). Set the values of the elements of the covariance matrix V.

ii) Give initial guess for $\underline{\theta}$

iii) Start computations for $i = 1, \ldots, n$: Calculate the elements of B_i, making $\hat{\underline{\xi}}_i = \underline{z}_i$. Calculate the value of Q_i as given by Eq. 8, which in this case reduces to

$$Q_i = \underline{g}_1(\underline{z}_i,\underline{\theta})' (\underline{B}_i \underline{V} \underline{B}_i')^{-1} \underline{g}(\underline{z}_i,\underline{\theta})$$

iv) Compute $\Sigma_{i=1}^{n} Q_i$. The objective is to minimize the value of this summation. On the basis of the optimization routine being used, decide next trial value for $\underline{\theta}$ and go back to iii). Termination criteria can be either stable values for $\underline{\theta}$ or for $\Sigma_{i=1}^{n} Q_i$ between two iterations.

Exact Solution. The steps i) and ii) will be identical to those described for the approximate solution.

iii) Start computations for $i = 1, \ldots, n$:

I. Calculate the elements of the matrix B_i, using the most recent estimate of $\hat{\underline{\xi}}_i$. (Only for the very first iteration this will be $\hat{\underline{\xi}}_i = \underline{z}_i$).

II. Find a new estimate $\hat{\underline{\xi}}_i^{(l)}$ by using Eq. 10.

III. Check for convergence on $\hat{\underline{\xi}}_i$. This can be done by calculating the value of Q_i using the current value of $\hat{\underline{\xi}}_i$ (say $\hat{\underline{\xi}}_i^{(l)}$) and comparing this to the value of Q_i at the new value (say $\hat{\underline{\xi}}_i^{(l+1)}$). If convergence has not been reached update the estimate of $\hat{\underline{\xi}}_i$ with the new value and go back to I. If convergence has been reached move to the next value of i, and go back to I. If convergence has been reached for the last value of i ($i = n$), go to the next step.

iv) Identical to step iv) of the approximate solution. As can be seen, the only difference between the exact and approximate solution is in step iii). The two solutions can be implemented in the same computer program, by allowing for the calculation of the estimate of $\hat{\underline{\xi}}_i$ if so desired.

LITERATURE CITED

Anderson, T. F., D. S. Abrams, and E. A. Grens, "Evaluation of Parameters for Nonlinear Thermodynamic Models," *AIChE J.*, **24**, 20 (1978).

Bard, Y., "Nonlinear Parameter Estimation," Academic Press, NY (1974).

Barker, J. A., "Determination of Activity Coefficients from Total Pressure Measurements," *Austr. J. Chem.*, **6**, 207 (1953).

Box, G. E. P., and D. R. Cox, "An Analysis of Transformations," *J. Royal Stat. Soc.*, (B), **26**, 211 (1964).

Box, G. E. P., and N. R. Draper, "The Bayesian Estimation of Common Parameters from Several Responses," *Biometrika*, **52**, 355 (1965).

Box, G. E. P., W. G. Hunter, J. F. MacGregor, and J. Erjavec, "Some Problems Associated with the Analysis of Multiresponse Data," *Technometrics*, **15**, 33 (1973).

Box, G. E. P., and G. C. Tiao, "Bayesian Inference in Statistical Analysis," Addison-Wesley Pub. Co., Don Mills, Ontario, Canada (1973).

Box, M. J., "Improved Parameter Estimation," *Technometrics*, **12(2)**, 219 (1970).

Brinkman, M. D., L. C. Tao, and J. H. Weber, "A Study of the Effect of the Optimization Function on Calculated Parameters of the Wilson Equation," *Can. J. Chem. Eng.*, **52**, 397 (1974).

Britt, H. I., and R. H. Luecke, "The Estimation of Parameters in Nonlinear Implicit Models," *Technometrics*, **15(2)**, 233 (1973).

Bruin, S., "Activity Coefficient Relations in Miscible and Partially Miscible Multicomponent Systems," *Ind. Eng. Chem. Fund.*, **9**, 305 (1970).

Deak, G., "Vapor-Liquid Equilibrium from Total Pressure Measurements," *Hung. J. Ind. Chem.*, **2**, 95 (1974).

Fredenslund, A., J. Gmehling, and P. Rasmussen, "Vapor-liquid equilibria using UNIFAC, a group contribution method," Elsevier, Amsterdam (1977).

Gmehling, J., and U. Onken, "Vapor-liquid equilibrium data collection," **1**, Part 1, Dechema, West Germany (1975).

Hayden, T., and J. O'Connel, "A Generalized Method for Predicting Second Virial Coefficients," *IEC Proc. Des. Dev.*, **14(3)**, 209 (1975).

Hirata, M., S. Ohe, and K. Nagahama, "Computer aided data book of vapor-liquid equilibria," Tokyo, Japan (1975).

Holmes, M. J., and M. VanWinkle, "Prediction of Ternary Vapor-Liquid Equilibria from Binary Data," *Ind. Eng. Chem.*, **62**, 21 (1970).

Kemeny, S., and J. Manczinger, "Treatment of Binary Vapor-Liquid Equilibrium Data," *Chem. Eng. Sci.*, **33**, 71 (1978).

Ma, K. T., C. McDermott, and S. Ellis, *I. Chem. Eng. Symp. Ser.*, **32(3)**, 104 (1969).

Marina, J. M., and D. P. Tassios, "Effective Local Composition in Phase Equilibrium Correlations," *Ind. Eng. Chem. Proc. Des. Dev.*, **12**, 67 (1973).

Patino-Leal, H., "Advances in the theory of the Error in Variables Model with applications in Chemical Engineering," Ph.D. Thesis, University of Waterloo (1979).

Prausnitz, J. M., C. A. Eckert, R. V. Orye, and J. P. O'Connell, "Computer calculations for multicomponent vapor-liquid equilibria," Prentice-Hall, Englewood Cliffs, NJ (1967).

Pritchard, E., T. Downie, and D. Bacon, "Further Consideration of Heteroscedasticity in Fitting Kinetic Models," *Technometrics*, **19(3)**, 227 (1977).

Reilly, P. M., "The Numerical Computation of Posterior Distributions in Bayesian Statistical Inference," *J. Royal Stat. Soc.*, (C), **25(3)**, 201 (1976).

Reilly, P. M., and H. Patino-Leal, "A Bayesian Study of the Error in Variables model," *Technometrics*, **23**(3) 221 (1981).

Sutton, T. L., and J. F. MacGregor, "The Analysis and Design of Binary Vapor-Liquid Equilibrium Experiments," *Can. J. Chem. Eng.*, **55**, 602 (1977).

Van Ness, H. C., S. M. Byer, and R. E. Gibbs, "Vapor-Liquid Equilibrium" *AIChE J.*, **19**, 2, 238 (1973).

Verhoeye, L. A. J., "Remarks on the Determination of the Wilson Constants in the Correlation of Vapor-Liquid Equilibrium Data," *Chem. Eng. Sci.*, **25**, 1903 (1970).

Wilson, A., and E. L. Simons, "Vapor-Liquid Equilibrium of Propanol-Water System," *Ind. Eng. Chem.*, **44**, 2214 (1952).

Yen, L. C., and S. S. Woods, "A Generalized Equation for the Computer Calculation of Liquid Densities," *AIChE J.*, **12(1)**, 95 (1966).

Manuscript received October 3, 1979; revision received August 24, and accepted August 31, 1981