CHEMICAL ENGINEERING

The numerical problems within analytical methods of solution for cubic equations of state

INGENIERÍA QUÍMICA

Los problemas numéricos dentro de los métodos analíticos de solución para ecuaciones de estado cúbicas

Javier I. Carrero-Mantilla[§]*

**Universidad Nacional de Colombia, Sede Manizales, Manizales-Colombia.*

[§] *jicarrerom@unal.edu.co*

(Recibido: Marzo 1 de 2011-Aceptado: Noviembre 21 de 2011-Version Final: Mayo 18 de 2012)

Abstract

It is possible to solve analytically cubic equations of state when they are reduced to a polynomial form, but it is known that at certain conditions application of the Cardano-Vieta formulas can produce wrong liquid density results due to numerical errors. In this work the same behavior was found in the hybrid analytical-iterative Deiters solution method, the causes of the errors were revisited, and for each method a new criterion was proposed to stop the calculation when wrong results can be produced. But it was also found that the wrong results can be avoided either using the reduced density as variable in the polynomial associated to the equation of state; or calculating the complete set of polynomial roots with the Jenkins-Traub algorithm, which can be even more advisable than any of the two aforementioned methods.

Keywords: Cubic equations of state, Cardano-Vieta method, Polynomial roots.

Resumen

Es posible resolver analíticamente ecuaciones cúbicas de estado cuando se reducen a una forma polinomial, pero se sabe que a ciertas condiciones la aplicación de las fórmulas de Cardano-Vieta puede producir resultados erróneos de densidad de líquido a ciertas condiciones debido a errores numéricos. En este trabajo se encontró el mismo comportamiento en el método de solución híbrido analítico-iterativo de Deiters a dichas condiciones. Las causas de los errores fueron reexaminadas, y para cada método se propuso un nuevo criterio para detener el cálculo cuando se pueden producir resultados erróneos. Pero también se encontró que los resultados erróneos se pueden evitar usando la densidad reducida como variable en el polinomio asociado a la ecuación de estado; o calculando el conjunto completo de raíces del polinomio con el algoritmo Jenkins-Traub, lo que puede ser incluso más recomendable que cualquiera de los dos métodos mencionados previamente.

Palabras Claves: Ecuaciones cúbicas de estado, Método Cardano-Vieta, Raíces de polinomios

1. Introduction

Cubic equations of state (cubic EOS) come from the addition of a covolume parameter (*b*) and an attractive pressure term inversely proportional to V^2 to the ideal gas equation. However, most modern cubic EOS use a polynomial of degree 2 instead of V^2 in the form

$$P = \frac{RT}{V-b} - \frac{a}{V(V+d)+c(V-d)}, \qquad (1)$$

where *P*, *V*, and *T* represent pressure, molar volume, and temperature; *R* is the gas constant; and the terms *a*, *b*, *c*, and *d* can be constant or functions of temperature and fluid properties including critical temperature and pressure, and acentric factor, Valderrama (2003). Any cubic EOS in the form of Eq. 1 can be rewritten as a cubic polynomial of the compressibility factor, defined as $Z=PV/(RT)$. For example the original van der Waals EOS (*c* =0and *d*=0 in Eq. 1, van der Waals (1873)) becomes

$$Z^3 - (1+B)Z^2 + AZ - AB = 0 \qquad (2)$$

with the dimensionless parameters $A= aP/(RT)^2$, $B=bP/(RT)$. Moreover, it results possible to write cubic EOS as polynomials in terms other than *Z*, such as the density reduced by the covolume, $\tilde{\rho}=b/V$, Deiters (2005), or a reduced molar volume defined as $\tilde{V}=b/V$. With these arrangements the van der Waals equation becomes

$$\tilde{\rho}^3 - \tilde{\rho}^2 + \tilde{\rho}\left[\frac{B}{A}(1+B)\right] - \frac{B^2}{A} = 0, \qquad (3)$$

or

$$\tilde{V}^3 - \tilde{V}^2\left(\frac{1+B}{B}\right) + \tilde{V}\left(\frac{A}{B^2}\right) - \frac{A}{B^2} = 0. \qquad (4)$$

The key to use a cubic EOS is the calculation of the density of the phases, necessary to obtain thermophysical properties, for example, residual enthalpy or fugacity coefficient. In cubic EOS pressure is written as the dependent variable (Eq. 1), therefore the calculation of pressure from temperature and volume, $P=P(T,V)$, is straightforward. However obtaining the molar volume from temperature and pressure, the $V=V(T,P)$ calculation, becomes a non-linear problem that requires a numerical solution and an initial estimate of the value of the root, or roots. Transforming the cubic EOS into a cubic polynomial of the form

$$x^3 + c_1x^2 + c_2x + c_3 = 0, \qquad (5)$$

with $x=Z$, $x=\tilde{\rho}$, or $x=\tilde{V}$ simplifies the non-linear problem because the fundamental theorem of algebra states that it has three roots, which can be found from three kinds of methods:

Analytic: the three roots are computed using the Cardano-Vieta (CV) formulas,Weisstein (2009, 2009b).

Iterative: an initial estimate of the value of the root is improved in successive steps using a numerical method. For real roots the Newton-Raphson, or Muller method can be applied. In the Deiters method the first root is found iteratively and used to deflate the polynomial to a quadratic form, Deiters (2002).

Multiple root finder (MRF): the complete set of roots is computed using the fact that they are eigenvalues of the companion matrix associated to the polynomial. Within the available techniques the Jenkins-Traub algorithm (also known as RPOLY) has become a *de facto* standard for numerical software, Jenkins (1975), Jenkins & Traub (1970, 1975), Press et al. (1992).

Cubic EOS have become a common model choice for process simulation and optimization due to their capability to predict properties of gas and liquid phases, therefore selection of the solution method for the polynomial roots in Eq. 5 is extremely important because being at the lowest levels in the execution hierarchy of the required iterative procedures (for example flash calculations) the functions related to the implementation of cubic EOS are frequently called. This implies that the overall computation time depends on their speed, and the reliability of the global result depends on their precision.

Solution of Eq. 5 with analytic or multiple root finder methods has two advantages over the use of iterative methods, initial estimates of the answer are not required, and the result is the complete set of three roots making possible to know directly if the EOS predicts densities for one or two phases from the number of real roots in the results. But it is necessary to consider also the speed and precision of the methods.

The relative speed of the different methods of solution depends strongly on the way that the algorithms and functions are programmed and has been discussed in previous works, Deiters (2002, 2005), Mathias & Benson (1986), Salim (2005). An informal test done for this work showed that using an optimized implementation the MRF method can be faster than the other ones; however the execution times for individual calculations were of milliseconds in all cases, even using interpreted languages. This suggests that speed is not a practical criterion to differentiate the methods of solution, as using modern processors the time required for complex simulations would increase or decrease in a few seconds due to the choice of a particular method. Hence in this work execution time was not considered to compare the different solution methods.

On the contrary, due to the disastrous effects that accumulated errors in the values of thermophysical properties could cause in a simulation, precision cannot be compromised for the calculation of the roots of a cubic EOS, and it is also necessary to check if the roots are physically feasible, Deiters (2002), Mathias et al. (1984). The analytic nature of the CV method suggests a guaranteed result, but its dependence on the functions square root, cos, and arccos can lead to a loss of numerical precision. In fact the CV formulas can fail in a catastrophic way producing physically unfeasible compressibility factors for the liquid phase due to an error magnification, as it was found by Zhi & Lee (2002) for some cubic EOS at a specific set of low temperature conditions. To avoid a possible failure of CV it has been proposed to simply avoid the analytical method at cryogenic conditions, Zare Nezhad & Eggeman (2006); however this strategy ignores that the values fed to the CV formulas, not the temperature, are the cause of the problem and thus similar failures could appear for other conditions of temperature and pressure with different cubic EOS. Hence a reliable solution procedure for Eq. 5 should have a safeguard against the use of the CV formulas based on numerical grounds, not on physical information.

In this work, the causes of the CV method failure and the criteria previously proposed to quit the CV calculation are revisited, and alternate methods of solution are analyzed in order to propose new numerical criteria to avoid that kind of failure. It was found that the error magnification proposed by Zhi & Lee (2002) can be discarded as the cause; that the combination of CV with an iterative procedure proposed by Deiters (2002) tends to fail at the same conditions reported for CV; and that the use of the $\tilde{\rho}$ as variable in Eq. 5 or the Jenkins-Traub algorithm prevents the numerical failure.

The remaining of this paper is organized as follows, in Methodology section the CV and Deiters solution methods are reviewed, it is explained how their results are tested, and the previous explanation about the failure of the CV method is discussed. In Results and discussion section, the numerical causes for the failure of the CV and Deiters methods are explained first, and finally, the criteria to avoid such failure are developed. Complete tables of results are available from the author as supplementary material.

2. Methodology

2.1 The Cardano-Vieta and Deiters methods

The analytical method of solution for cubic polynomials is based on the Vieta substitution within the formulas of Cardano, Weisstein (2009, 2009b), as described in Fig. 1. This algorithm starts with the calculation of the polynomial discriminant, *d*, defined as

$$d = r^2 + q^3, \qquad (6)$$

with

$$q = \frac{3c_2 - c_1^2}{9}, \qquad (7)$$

and
$$r=\frac{c_1c_2}{6}-\frac{c_3}{2}-\frac{c_1^3}{27}. \qquad (8)$$

The number of real roots of Eq. 5 depends on *d*, if d>0 there is only one real root (calculated as $s+t-c_1/3$ with $s=\sqrt[3]{r+\sqrt{d}}$ and $t=\sqrt[3]{r-\sqrt{d}}$) and the other two are complex conjugate, otherwise there are three real roots that come from the expression

$$x=2\sqrt{-q}\cos\left(\frac{\theta+2\pi j}{3}\right)-\frac{c_1}{3}. \qquad (9)$$

with *j*=0,1,2. The angle θ (in radians) is calculated as

$$\theta=\arccos(s), \qquad (10)$$

where

$$s=\frac{r}{\sqrt{-q^3}}. \qquad (11)$$

In the Deiters method (see Deiters (2002), Fig. 2) the cubic polynomial is defined as the $g(x)$ function

$$g(x)=x^3+c_1x^2+c_2x+c_3, \qquad (12)$$

whose real roots are within the interval $[-\rho,\rho]$, where

$$\rho=1+\max(|c_1|,|c_2|,|c_3|) \qquad (13)$$

The first root is found iteratively and the initial estimate is one of the two extremes of the interval, assigned according to

$$x^{(0)}=\begin{cases}-\rho & \text{if} \quad g(x_{\text{ifl}})>0\\ +\rho & \text{if} \quad g(x_{\text{ifl}})\le 0\end{cases} \qquad (14)$$

where $x_{\text{ifl}}=-1/3c_1$ is an inflection point of $g(x)$.Then the first and second derivatives of *g* are used with the Halley's method, also named "Kepler's method" by Deiters,

$$x^{(k+1)}=x^{(k)}-\left[\frac{gg'}{(g')^2-gg''/2}\right]^{(k)} \qquad (15)$$

until the value of the *k*-th estimation of the root converges to a value x_1, Koçak (2008, 2008a), Weisstein (2009a).The two remaining values of *x* are the roots of the quadratic polynomial

$$h(x)=x^2+bx+c \qquad (16)$$

with coefficients $b=x_1+c_1$ and $c=bx_1+c_2$, obtained applying polynomial deflation to $g(x)$. In this way the Deiters method does not require an initial estimate of the roots and combines the advantages of the numerical iterative and CV methods.

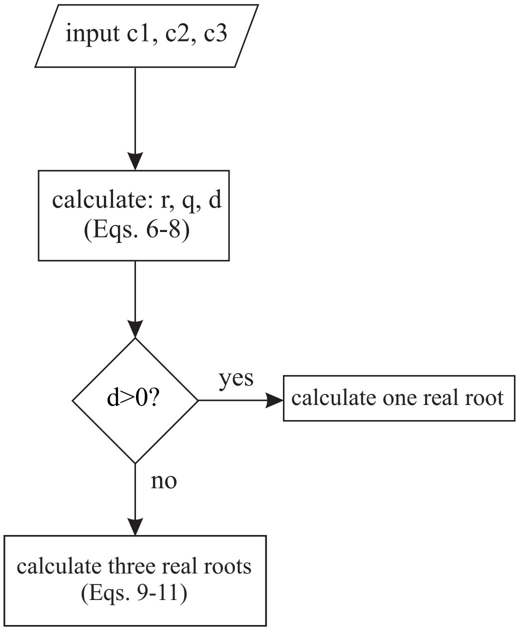

Figure 1: Flowchart describing the solution algorithm for the Cardano-Vieta method.

2.2 The tests for the solution methods

In a previous work it was found that for extremely low pressures and temperatures liquid volume (V_L) results from cubic EOS calculations depend on the method of solution, Zhi & Lee (2002). Results from the application of the Newton-Raphson (NR) method with *x*=*Z* in Eq. 5 had a very acceptable

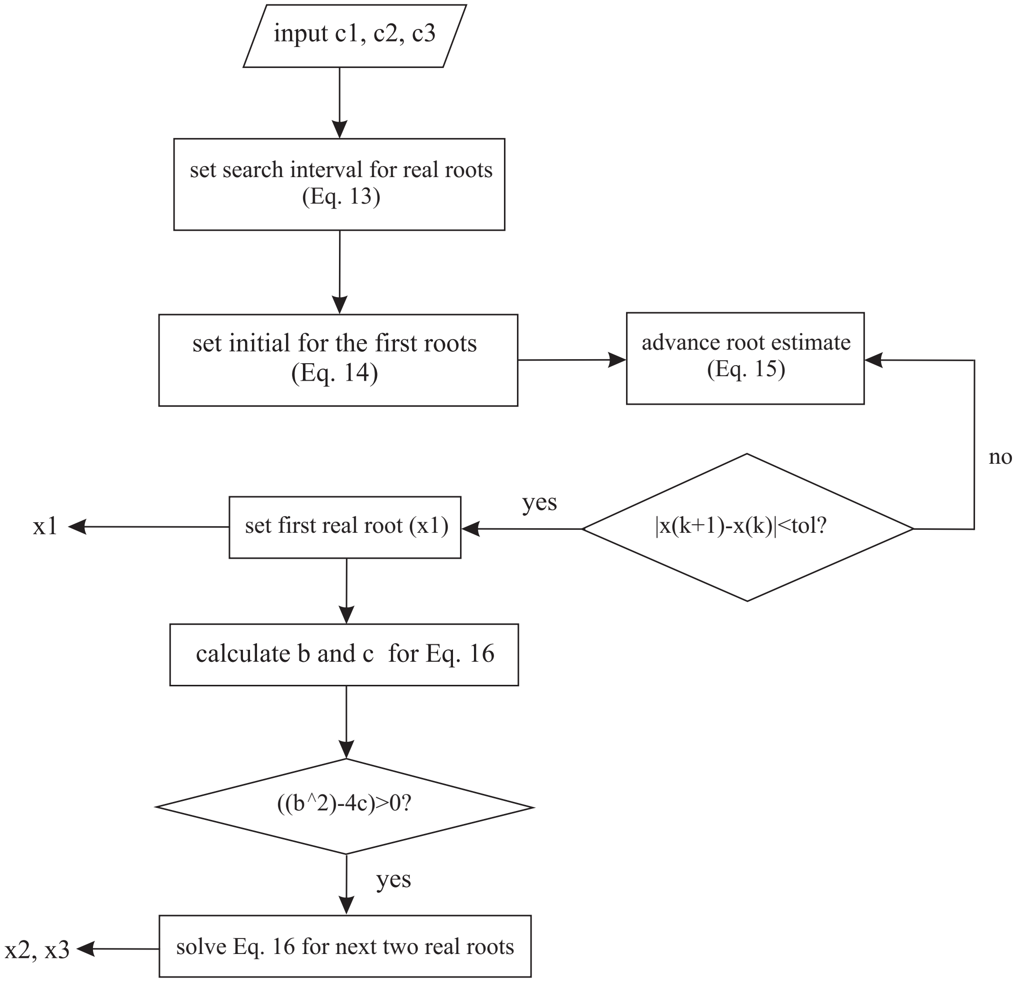
input c1, c2, c3
set search interval for real roots (Eq. 13)
set initial for the first roots (Eq. 14)
advance root estimate (Eq. 15)
no
|x(k+1)-x(k)|<tol?
yes
x1
set first real root (x1)
calculate b and c for Eq. 16
((b^2)-4c)>0?
yes
x2, x3
solve Eq. 16 for next two real roots

Figure 2: Flowchart describing the algorithm for the Deiters method.

agreement with experimental values while CV results were wrong, being in some cases physically infeasible. Considering that the use of equations of state to predict liquid densities has been deemed not advisable due to deviations of the results respect to the experimental values, Valderrama & Alfaro (2000), the production of unreliable V_L predictions only in a very specific T,P region could be attributed to limitations in the range of application of the cubic EOS but it is not the case: if the failures were attributable to the cubic EOS both CV and NR results should coincide to the same wrong values, i.e. results from both methods of solution should diverge from experiment in the same way. Moreover, with the three cubic EOS used by Zhi and Lee reported saturated liquid density results tend to remain within the same order of magnitude of the experimental values and are positive: reported discrepancies for various substances and conditions are <30% for the cubic EOS Peng-Robinson (PR, Peng & Robinson (1976)), <5% for the cubic EOS Chain of Rotators (CCOR, Kim et. al. (1986), Lin (1983)), and <7% for Patel-Teja (PT, Teja & Patel (1982)). Instead, in the worst results reported by Zhi and Lee there are differences of several orders of magnitude respect to the experimental values, or even negative values that cannot be attributed to errors in the implementation of the CV method, as it produces results coincident with NR in other regions.

The failure of the CV formulas has been attributed to numerical inconsistencies which are revisited in this work not only for CV but also for the Deiters method using the same benchmark of 48 conditions and three cubic EOS proposed by Zhi & Lee (2002), including temperatures and pressures in the ranges 87.8K≤*T*≤266.3K and 3.56×10^{-7} Pa≤*P*≤1.14×10^{5} Pa for propylene, 1-butene, and 1-pentene. For each *PT*, condition results were obtained with the PR, CCOR, and PT cubic EOS using $x=Z$, $x=\tilde{\rho}$, and $x=\tilde{V}$ as variables in Eq. 5. The CV, and Deiters methods were written for the FORTRAN compiler Absoft (2008); and the Jenkins-Traub algorithm, as implemented within the function roots included in the software Scilab (2009), was used as MRF. Critical properties and acentric factors were taken from Perry's handbook, Liley et al. (1999). As happened with the NR method in the original Zhi and Lee study it was found that V_L values from the MRF used agree with the experimental ones in the whole range of conditions (see Table 1). This indicates that MRF does not fail at the *T*, *P* conditions of the benchmark, hence *Z*, $\tilde{\rho}$, and $\tilde{V}$ results from CV and Deiters methods were compared against the values from MRF.

The liquid phase specific volume, V_L, was obtained from the corresponding root of Eq. 5 (with $x=Z$) as

$$V_L = Z_L \frac{RT}{P}, \qquad (17)$$

but the comparison with experimental V_L values was avoided because in this case the agreement depends not only on the solution method but also on the accuracy of the EOS, which is not the object of study of this work. For *Z* and $\tilde{V}$ the differences between CV and Deiters, and MRF results were measured in a base-10 logarithmic scale defining the parameters

$$p_{ZL} = |1\text{n}\,(Z_L^n) - 1\text{n}\,(Z_L)|/1\text{n}\,(10), \qquad (18)$$

and

$$p_{VL} = |1\text{n}\,(\tilde{V}_L^n) - 1\text{n}\,(\tilde{V}_L)|/1\text{n}\,(10), \qquad (19)$$

where the superscript *n* identifies the values from the MRF method.

Table 1: Molar volumes of liquid propylene, in mol/cm3*, obtained with the Z form (x=Z in Eq. 5).* EXP *represents the experimental values,* PT *results from the Patel-Teja EOS,* PR *results from the Peng-Robinson EOS, and* CC *results from the Cubic Chain of Rotators EOS. Results for 1-butene and 1-pentene are available as suplementary material from the author.*

			MRF			Cardano-Vieta			Deiters		
T/K	*P*/MPa	EXP	PT	PR	CC	PT	PR	CC	PT	PR	CC
87.9	9.18E-10	55.07	56.63	53.84	55.08	5.46E4	7.96E11	6.13E4	7.96E11	7.96E11	2.63E2
89.4	1.60E-9	55.18	56.70	53.90	55.22	3.23E4	3.22E4	3.84E4	4.65E11	4.65E11	1.92E3
90.9	2.74E-9	55.29	56.77	53.96	55.35	1.95E4	2.06E4	2.55E4	1.03E2	3.35E3	7.01E3
92.4	4.59E-9	55.39	56.85	54.03	55.49	1.30E4	1.29E4	1.79E4	4.40E2	77.75	6.76E3
93.9	7.56E-9	55.5	56.92	54.10	55.63	8.66E3	8.80E3	1.34E4	3.03E2	2.07E3	65.64
95.4	1.22E-8	55.61	56.99	54.16	55.77	6.16E3	6.15E3	1.06E4	1.98E3	1.99E3	6.26E3
96.9	1.95E-8	55.72	57.06	54.23	55.90	4.52E3	4.49E3	8.76E3	36.31	1.76E3	83.06
98.4	3.06E-8	55.84	57.14	54.30	56.04	3.50E3	3.48E3	7.58E3	26.28	81.59	51.84
99.9	4.73E-8	55.95	57.21	54.37	56.18	2.84E3	2.82E3	6.76E3	50.39	45.63	54.29
101.4	7.21E-8	56.06	57.29	54.44	56.31	2.41E3	2.39E3	6.18E3	56.33	52.97	57.67
102.9	1.08E-7	56.17	57.37	54.51	56.45	2.13E3	2.10E3	5.74E3	57.49	53.97	56.88
127.9	2.08E-5	58.17	58.77	55.80	58.69	57.05	54.08	56.97	58.77	55.80	58.69
152.9	6.03E-4	60.39	60.46	57.35	60.95	60.39	57.28	60.88	60.46	57.35	60.95
177.9	6.09E-3	62.91	62.50	59.22	63.32	62.49	59.22	63.31	62.50	59.22	63.32
202.9	3.24E-02	65.79	65.01	61.55	65.91	65.01	61.54	65.91	65.01	61.55	65.91
227.9	1.14E-01	69.16	68.18	64.48	68.88	68.18	64.48	68.88	68.18	64.48	68.88

2.3. Analysis of the behavior of the Cardano-Vieta formulas

Zhi and Lee (2002) defined

$$M_c = \frac{\Delta Z_L / Z_L}{\Delta s / s}, \qquad (20)$$

as an "error magnification" in the CV formulas, where Δ indicates the absolute error. For the liquid root, *j*=1 in Eq. 11, and approximating Δ to the differential M_c becomes

$$M_c = \frac{s}{Z_L}\left(\frac{dZ_L}{ds}\right) \qquad (21)$$

$$= \frac{2\sqrt{-q}}{3Z_L}\sin\left(\frac{\arccos(s)+2\pi}{3}\right)\frac{s}{\sqrt{1-s^2}}.$$

Given that

$$\lim_{s\to 1}\frac{s}{\sqrt{1-s^2}} \to \infty, \qquad (22)$$

M_c tends to infinity when $s \to 1$ and induces a huge error in Z_L, hence the failure of CV was attributed to the condition $s \to 1$.

In this work the magnification error was constrained to finite values replacing the function arccos with arctan to calculate the roots from

$$Z = 2\sqrt{-q}\cos\left(\frac{\arctan(\sigma)+2\pi j}{3}\right) - \frac{c_1}{3} \qquad (23)$$

with *j*=0,1,2. This form correspond to the same angle θ in Eq. 9 with

$$\sigma = \sqrt{\frac{-q^3}{r^2} - 1}. \qquad (24)$$

Using Eq. 23 the error magnification for the liquid root becomes

$$M_t = \frac{\sigma}{Z_L}\left(\frac{dZ_L}{d\sigma}\right) \qquad (25)$$

$$= -\frac{2\sqrt{-q}}{3Z_L}\sin\left(\frac{\arctan(\sigma)+2\pi}{3}\right)\frac{\sigma}{1+\sigma^2}$$

where the subscript *t* indicates the use of arctan instead of arccos. The value of M_t must remain finite and close to zero when 1→*s* for two reasons; first given that

$$\lim_{s\to 1}\left[\frac{\sigma}{1+\sigma^2}\right] = 0, \qquad (26)$$

and second, $|M_t| = |M_c|(1-s^2)$, therefore M_t will tend to zero when *s*→1 as long as M_c remains finite. The proximity of *s* to 1 was measured using the parameter

$$p_s = -\ln(1-s)/\ln(10), \qquad (27)$$

for example $s = 0.99 \Rightarrow p_s = 2$, $s = 0.9999 \Rightarrow p_s = 4$ and so on. Results in Fig. 3 show, as expected, that M_t falls below M_c for all conditions, even when *s*→1. But despite the reduction of *M* value by several orders of magnitude the differences between Z_L from CV and MRF when *s*→1 remain in the results (Fig. 4). This shows that the "magnification error" is not the origin of the apparent failure of CV method, and it makes necessary a different analysis of the CV formulas.

3. Results and discussion

3.1 Error propagation in the Cardano-Vieta method

Complete agreement with the experimental V_L values was found for all conditions and EOS if the polynomial in Eq. 5 is used with $x=\tilde{\rho}$, or if the MRF method is used with any of the three options, $x=Z$, $x=\tilde{\rho}$, or $x=\tilde{V}$. On the other hand, for the lowest *T* values in the benchmark application of CV or Deiters methods produced completely wrong V_L values. Certainly these results are induced by the values of and *P* used as arguments for the EOS, but their direct cause is the application of the CV and Deiters formulas to polynomials with coefficients that differ by several orders of magnitude, which induce numerical truncation errors that propagate to the final results. In order to show how the propagation takes place the calculation of the three real roots for propylene at 95.4 K and $1.22\times10^{\wedge-2}$Pa with the Patel-Teja EOS, Teja& Patel (1982), is used as an illustrative

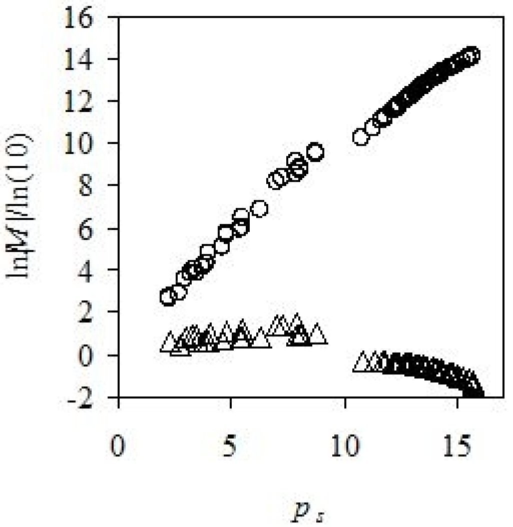

Figure 3: Magnification factors obtained using arccos (M_c, circles) and arctan (M_t, triangles) functions as a function of the p_s parameter. M_c and M_t are represented in the vertical axis with the base-10 logarithms of their absolute values. The limit s→1 corresponds to the higher values of p_s.

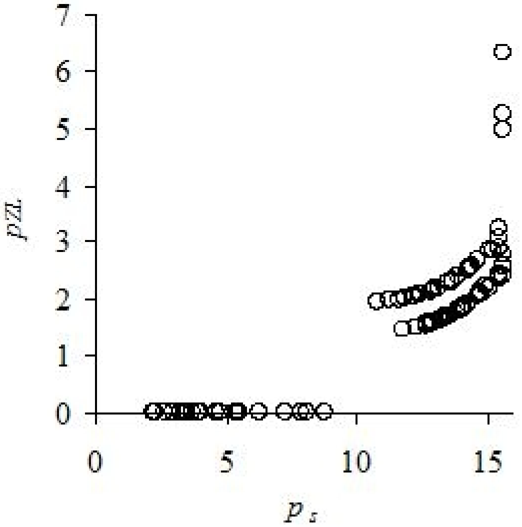

Figure 4: Logarithmic difference between Z_L from Cardano-Vieta, using arctan in Eq. 23, and MRF methods as function of p_s.

example (see Table 2). For $x=Z$ in Eq. 5 there are differences of eight and nine orders of magnitude between c_1 and c_2 and between c_2 and c_3 respectively, in the calculation of q, r, and s these differences propagate through the operations. In Eq. 7, $q=(3c_2-c_1^2)/9$ the terms, $c_1^2=$ 0.999999999126088 and $3c_2=8.41327018500574\times10^{-8}$ have a difference of eight orders of magnitude implying that most of the digits in the $3c_2$ term do not count for the subtraction, as typically the 15 leftmost digits are significant with the DOUBLE PRECISION variables, Compaq (2000). Next, in Eq. 8 again numbers of very different orders of magnitude are being subtracted in the calculation of the term $r=c_1c_2/6-c_3/2-c_1^3/27$ with $c_1c_2/6=-4.6740389896275\times10^{-9}$, $c_3/2=-1.19069048757051\times10^{-17}$, and $c_1^3/27=-3.70370369884864\times10^{-2}$, this means that most of the digits of $c_3/2$ are lost when it is subtracted from $c_1c_2/6$, and it is practically null when compared with $c_1^3/27$. The round-off errors in the calculation of q and r are propagated to s through Eq. 11, $s=r/\sqrt{-q^3}$ and s→1 (p_s =14.6) because the numerator and denominator differ in a few digits ($\left|r-\sqrt{-q^3}\right|=9.02\times10^{-17}$) next, the calculation of θ using s in Eq. 10 amplifies the original round-off errors due to the use of the arccos function and with this erroneous argument, the cosine function in Eq. 9 produces wrong Z results. Defining the function

$$p(x)=\ln|x|/\ln(10), \qquad (28)$$

to measure the order of magnitude of the terms $x=c_1$^2, $x=3c_2$, $x=c_1c_2/6$, $x=c_3/2$, and $x=c_1^3/27$ it is shown in Figure 5 that these huge differences between the terms used to calculate q and r appear in all cases when s→1.

This error propagation occurs for the three Z roots, but its effects are noticeable only when $|Z|\rightarrow0$, as it happens with Z_L (j=1 in Eq. 9). In the example MRF results for Z are $\{1.00, 2.72\text{x}10^{-8}, 8.77\text{x}10^{-10}\}$, while CV gives $\{1.00, -3.31\text{x}10^{-8}, 9.48\text{x}10^{-10}\}$. The Z=1.00 root is not changed because the error propagation affects the last decimal places, while the other two, both close to zero, are dramatically affected. This suggests that Z results close to zero from the CV method must not be trusted, even using double precision variables.

The order of magnitude of the coefficients changes if $x=\widetilde{V}$ is used in Eq. 5. For the same

Table 2: Coefficients $_ic$ *in Eq. 5 and intermediate parameters in the Cardano-Vieta and Deiters methods for propylene at* 95.4K *and* $1.22\times10^{\wedge-2}$Pa *with the Patel-Teja EOS. xrepresents the variable, compressibility factor, reduced density, or reduced volume and* x_L *is the result corresponding to the liquid. The function p is defined in Eq. 28.*

x =	Z	$\tilde{\rho}$	V
x_L (MRF)	8.76549101693965×10	0.941063756188202	1.0626272592311
x_L (CV)	$9.47710495635689\times10^{-8}$	0.9410637546022087	115.9657862186432
c_1	-0.9999999995630439	-0.9714266137223527	$-1.212284923269059\times10^{9}$
c_2	$2.804423395001912\times10^{-8}$	$2.85733862291067\times10^{-2}$	$4.121478037063378\times10^{10}$
c_3	$-2.381380975141026\times10^{-17}$	$-2.356986025368972\times10^{-11}$	$-4.242706529596227\times10^{10}$
$p\ (c_1^2)$	-3.795×10^{-10}	-2.518×10^{-2}	18.167
$p\ (3c^2)$	-7.075	-1.067	11.092
$p\ (c_1c_2/6)$	-8.330	-2.335	18.920
$p\ (c_3/2)$	-16.924	-10.929	10.327
$p\ (c_1^3/27)$	-1.431	-1.469	25.819
q (Eq.7)	$1.11111101665931\,8\times10^{-1}$	$9.53277230178619\times10^{-2}$	$-1.632927346156809\times10^{17}$
r (Eq.8)	$3.703703231444739\times10^{-2}$	$2.932590974309058\times10^{-2}$	$6.598577064220757\times10^{25}$
s (Eq.11)	0.9999999999999975	0.9963746374492553	0.9999999999999972
p_s (Eq.27)	14.612	2.441	14.557
b (Eq.16)	$-2.80442348232767\times10^{-8}$	$-3.036285912014403\times10^{-2}$	$-3.399760293960571\times10^{1}$
c (Eq.16)	$7.45244534761608\times10^{-17}$	$2.504604509900865\times10^{-11}$	$5.549893188476562\times10^{1}$
$(b/2)^2$	$1.96619776705772\times10^{-16}$	$2.304758034874283\times10^{-4}$	$2.889592514097718\times10^{2}$
p_{bc}	-15.567	-3.637	2.368

example, in absolute value, c_1 becomes the smallest one and the exponent sign of c_2 and c_3 changes from negative to positive but anyway the liquid result is wrong (see Table 2). The huge difference between the terms used to calculate q and r produced p_s=14.557, inducing the failure of the CV procedure. This behavior appears again for all cases where p_s is large.

The apparent immunity to the failure of the calculation with $x=\tilde{\rho}$ in Eq. 5 is explained in a similar way. In the example the coefficient c_3 is several orders of magnitude lower than c_1 and c_2. This difference produces s = 0.996 but it is not close enough to 1 as to trigger the amplification of the round-off errors found for x=Z and x=$\tilde{V}$. In the calculation of q there is a significant subtraction of digits, and the same stands to a lesser degree for the calculation of r (see Table 2). This behavior appears in the rest of the benchmark, supporting the use of ρ instead of $\tilde{V}$ as recommended by Deiters (2005).

At this point, it is necessary to mention that the differences in order of magnitude that lead to numerical loss of precision were noticed in the

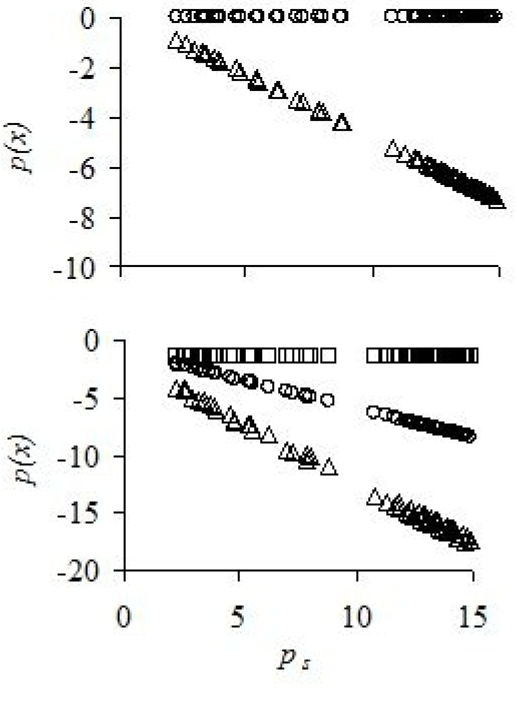

Figure 5: p-functions for the terms used to calculate q in Eq. 7 as function of p_s*. Upper half: circles are for with* $x=(c_1)^2$*, and triangles are for* $x=3c_2$*. Lower half: Circles are for* $p\,(x)$ *with* $x=c_1c_2/6$*, triangles are for* $x=c_3/2$*, and squares for* $x=c_1^3/27$*.*

work of Zhi and Lee, at least for the calculation of *r* using the *Z* form, but its effect when $s \to 1$ was explained with the "magnification error". Also, given that the failure comes from numerical effects and therefore is not directly caused by the kind of cubic EOS or substance the use of results from three EOS and three substances is reiterative: results from a single EOS and one substance would have been enough, but anyway the original set was maintained here for the sake of the continuity.

3.2 Error propagation in the Deiters method

Results from the application of the Deiters method were also wrong at the same conditions reported for the CV formulas. For the same illustrative example (Patel-Teja EOS applied to propylene at 95.4K and 1.22×10^{-2}Pa) the Deiters method produces $Z=\{1.00, 3.05\text{x}10^{-8}, -2.44\text{x}10^{-9}\}$ while the MRF result is $Z=\{1.00, 2.72\text{x}10^{-8}, 8.77\text{x}10^{-10}\}$. This implies that the sets of polynomial coefficients inducing the failure of CV formulas have the same effect on the Deiters method, although different numerical anomalies account for such effect.

The causes of the errors were traced back to the application of Eqs. (12)-(16), using the *Z* polynomial form it was observed for the failure conditions tested that $c_1 \approx -1$ and the first root, obtained iteratively, is $Z_1 \approx 1$. In the deflation of the cubic polynomial both *b* and *c* parameters tend to zero because $b=Z_1+c_1\approx0$, $|c_2|<<|c_1|$, and $c_2\approx0$, therefore $c=bZ_1+c_2\approx0$. When *b* and *c* are replaced in the solution of the quadratic polynomial in Eq. 16, $x=-b/2\pm\sqrt{(b/2)^2-c}$, the argument of the square root becomes very close to $\varepsilon=2.220\times10^{-16}$, the smallest possible value in double precision variables. For the example $(b/2)^2-c=2.7\times10^{-16}$ (Table 2), hence the round-off errors become comparable with the results and propagate through the final result. By contrast, when the method is applied to the polynomial $\tilde{\rho}$ form the *c* term comes close to zero but *b* does not, and the condition $|(b/2)^2-c|\to0$ is not reached, in the example $(b/2)^2-c=2.305\times10^{-4}$.

The reasons for the failure when the $\tilde{V}$ form is used are similar, since the variable has changed from *Z* to $\tilde{V}$ the first root and c_1 are no longer close to 1 but anyway they are almost equal and in their subtraction many of the significant digits are not used, for the example $\tilde{V}_1=1.21228488927145\times10^9$ and $c_1=-1.212284923269059\times10^9$ produces $b=\tilde{V}+c_1=-0.000000033997609\times10^9=-33.997609$. The *c* parameter comes from $c=bV_1+c_2$ and again it implies the subtraction of almost equal quantities, given that $b\tilde{V}_1\approx-4.12\times10^{10}$ and $c_2\approx4.12\times10^{10}$. Propagation of the round-off errors through these two operations cause the method to fail, even though the absolute value of $(b/2)^2-c$ is far from zero (233.46 in the example).

3.3 Strategies to avoid the error propagation

The errors induced by the representation of the numbers with a limited number of digits would be avoided using arbitrary-precision arithmetic, but this approach requires rewriting the code using specialized libraries. Instead the orders of magnitude of the three coefficients could be normalized using a scaling factor λ selected from

$$\lambda = \max\left(\sqrt[3]{|c_3|}, \sqrt{|c_2|}, |c_1|\right) \qquad (29)$$

to rewrite the polynomial in Eq. 5 as

$$y^3+\frac{c_1}{\lambda}y^2+\frac{c_2}{\lambda^2}y+\frac{c_3}{\lambda^3}=0 \qquad (30)$$

where $y=x/\lambda$, Deiters (2005). But for most of the conditions that induce the failure of the CV method with the *Z* form the scaling is negligible because $\lambda=|c_1|\approx1$. Therefore, the definition of λ was changed to scale the polynomial according to the orders of magnitude of the coefficients. With the new definition λ is generated from the term with the highest or the lowest order of magnitude choosing the c_i whose corresponding $|\log|c_i||$ is the largest, if the chosen coefficient is c_3 then $\lambda=\sqrt[3]{|c_3|}$, if c_2 then $\lambda=\sqrt{|c_2|}$ and if c_1 then $\lambda=|c_1|$. However, it was found that this scaling does not solve the numerical issues neither for the *Z* nor the $\tilde{V}$ form, it may only reverse the order of magnitude differences. For the example cited in Table 2 the coefficients (with the *Z* form) were changed from $\{-1.00, 2.80\text{x}10^{-8}, -2.38\text{x}10^{-17}\}$ to $\{3.48\text{x}10^{5}, 3.39\text{x}10^{3}, -1\}$ but anyway $s \rightarrow 1$ and the polynomial roots were the same $\{-3.31\text{x}10^{-8}, 9.48\text{x}10^{-8}\}$ and a similar behavior was observed in all related cases.

Rather than trying to change the polynomial coefficients it is preferable to define a criterion to turn off the CV procedure and switch to a numerical method before wrong *Z* or $\widetilde{V}$ values are produced, for example when T_r is below 0.3, Zhi & Lee (2002). The simplicity of this criterion makes it attractive, and it was found in this work that it also stands for the Deiters method applied to Z_L and $\widetilde{V}_L$, see Fig. 6. However it only results practical to obtain a mixture's density as a whole if pseudocritical properties are being used; otherwise, each component has its own critical temperature and the concept of T_r does not apply for the mixture, for example, when a mixing rule based on an activity coefficient model is being used, Orbey & Sandler (1998), Wong et al. (1992). A second objection comes from the fact that the $T_r \leq 0.3$ criterion depends on the critical temperature, while the failure of the CV formulas has numerical causes, not related with any physical property. This means that it cannot be guaranteed that *only* conditions with $T_r \leq 0.3$ will induce the failure, as other sets of substances, temperatures, and cubic EOS could produce polynomial coefficients of very different orders of magnitude in Eq. 5. A truly general criterion has to be based only on the parameters involved in the solution of the polynomial; neither the cubic EOS nor the physical properties should be involved.

It has been proposed another criterion, exclusively numerical and based on the behavior of the results for the Zhi and Lee's benchmark, to avoid the CV method if the Z_L value is below a "transition point" of 2×10^{-6}, Salim (2006). This condition indeed coincides with the e $T_r \leq 0.3$ criterion, as can be observed in the tables of results, and is reassured by the observation in Sec. 3.1 that *Z* roots close to zero are dubious. In such case, given that $Z_L^3 \approx 0$ the first term of Eq. 5 (with $x=Z$) vanishes and this equation gets reduced to a quadratic polynomial. But applying this criterion requires obtaining first Z_L from Eq. 5, either with the CV method, or with a numerical method.

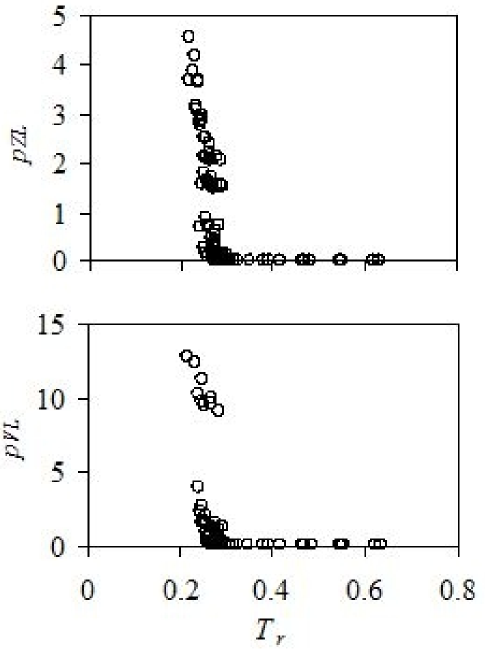

Figure 6: Logarithmic difference between Z_L (upper half) and $\widetilde{V}$ (lower half) from Deiters and numerical (MRF) methods as function of the reduced temperature.

Instead of checking the results from the CV method it is preferable to use a criterion based on the value of *s* as a "symptom" of the numerical failure, associated with the condition $s \rightarrow 1$. Defining the failure of CV as a difference of more than one order of magnitude respect to the MRF result, that is, $p_{ZL} > 1$, the plots of p_{ZL} and p_{VL} as functions of p_s in Fig. 7 show that such failure occurs when $p_s \geq 8$. Therefore it is recommended to interrupt the CV calculation and switch to an iterative or MRF method when $p_s \geq 8$.

Finally it is also possible to build a similar criterion for the Deiters method, the proximity of c_1 to the first root is measured in logarithmic scale defining the parameter

$$p_{c1x1} = \ln|x_1 - c_1| / \ln(10), \qquad (31)$$

which was applied to the results with $x=Z$ and $x=\widetilde{V}$. The results for the agreement between Deiters and MRF results in Fig. 8 suggest that the Deiters method should be avoided when $p_{c1Z1} < -6$ or $p_{c1V1} > 1.5$, which are the suggested conditions to stop the calculation.

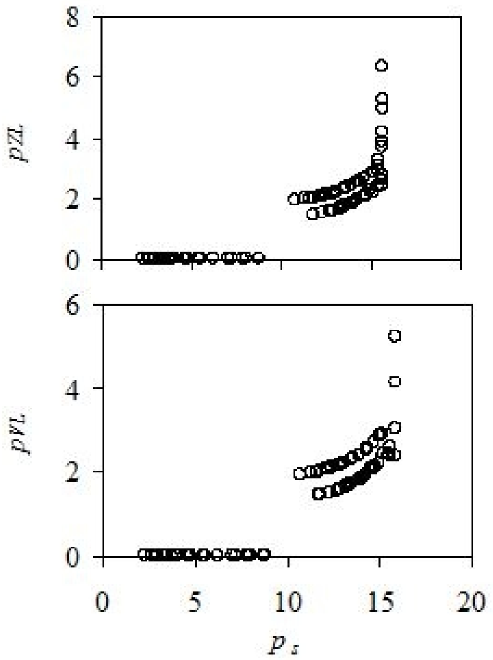

Figure7: Logarithmic difference between Z_L(upper half) and $\tilde{V}$ (lower half) from Cardano-Vieta and numerical (MRF) methods as function of p_s.

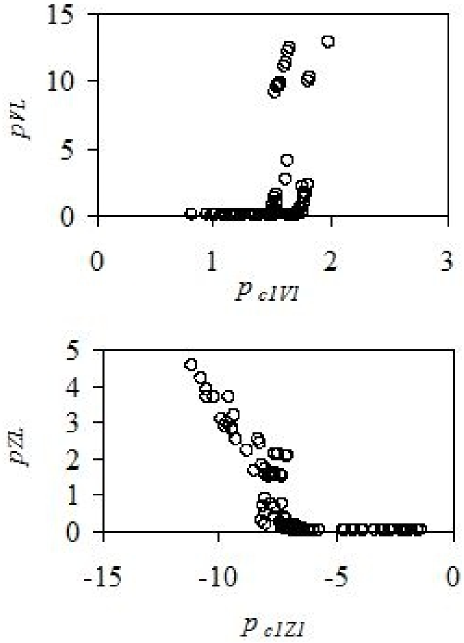

Figure 8: Logarithmic difference between Z_L (left) and $\tilde{V}$ (right) from Deiters and numerical (MRF) methods as function of p_{c1Z1} and p_{V1c1}.

4. Conclusions

The failure of analytic or semi-analytic methods of solution (Cardano-Vieta, Deiters) can be avoided in two different ways, casting the cubic EOS as a polynomial in terms of the reduced density, $\tilde{\rho}$, or using a numerical method of solution. In particular the Jenkins-Traub algorithm can be considered the numerical method of choice, since it has the advantage of providing all the roots of the polynomial in a single calculation, while purely iterative methods may require a different search for each root.

If the much more common Z, or even the $\tilde{V}$, polynomial form is to be solved with the CV formulas or the Deiters method it is possible to predict the failure before computing the roots. It was found that the CV procedure should be stopped if the parameter p_s is greater than a limiting value, namely if $p_s \geq 8$ with the polynomial in terms of Z or $\tilde{V}$. In the same way the Deiters method should be stopped if $p_{c1Z1} < -6$ with the polynomial in terms of Z, or $p_{c1V1} > 1.5$ with the polynomial in terms of $\tilde{V}$. Nevertheless these criteria must not be interpreted as strict constraints, other values may be chosen, for example a more conservative condition of $p_s \geq 7$. A previously proposed criterion based on the reduced temperature was found impractical, as in most of the cases it is not suitable for mixtures. In contrast, being at the core of the calculations, the proposed criteria are strictly numerical and general. They do not depend on the properties of the substance or mixture and can be applied to any cubic EOS model requiring only the additional calculation of a logarithm, saving computation time. But it is also necessary to consider that the Jenkins-Traub MRF is apparently immune to the failures induced by the polynomial coefficients, and with an optimized implementation it can be as fast as the analytical method. If it is possible to incorporate it to the rest of the simulation code, it would a better choice than CV, Deiters, or any iterative method of solution for the cubic polynomial.

As a final comment, the apparent failure of analytical formulas shows the perils of conceiving computational tools as black boxes that produce perfect results. Far from it, every operation performed on real numbers represented in binary format with a limited number of digits leads to a loss of accuracy for the result that in some strange cases can be amplified with catastrophic consequences for the final output.

Nomenclature

a	Cubic EOS parameter	Pa m^6 mol^{-2}
A	Dimensionless form of a	-
b	Covolume	m^3 mol^{-1}
B	Dimensionless form of b	-
c	Volume term in cubic EOS	m^3 mol^{-1}
c	Intermediate parameter in Deiters method	-
c_i	Polynomial coefficients (i=1, 2, 3)	-
d	Volume term in cubic EOS	m^3 mol^{-1}
d	Intermediate parameter in Cardano-Vieta method	-
M_c	Magnification factor from arccos	-
M_t	Magnification factor from arctan	-
P	Pressure	Pa
q	Intermediate parameter in CV method	-
r	Intermediate parameter in CV method	-
R	Gas constant, $R = 8.31451$	Pa m^3/(mol·K)
s	Intermediate parameter in CV method	-
T	Temperature	K
V	Molar volume	m^3 mol^{-1}
x	Polynomial variable	-
Z	Compressibility factor	-
ε	Smallest possible real value	-
θ	Angle used in the CV method	-
λ	Scaling factor	-
σ	Angle used in the CV method,	-
	Al p parameters are in base-10 logarithmic scale	
p_{bc}	Difference between $(b/2)^2$ and c (Deiters method)	-
p_{cij}	Difference between c_i and c_j	-
p_s	Proximity of parameter s to 1	-
$p(x)$	Logarithm of x	-
p_{ZL}	Difference between Z_L from CV (or Deiters) and MRF methods	-
p_{C1Z1}	Difference between c_1 and Z_1 (Deiters method)	-
p_{C1V1}	Difference between c_1 and $\widetilde{V}_1$ (Deiters method)	-
	Abbreviations	
CCOR	Cubic Chain Of Rotators equation of state	
CV	Cardano-Vieta	
EOS	Equation Of State	
MRF	Multiple Root Finder	
NR	Newton-Raphson method	
PR	Peng Robinson EOS	
PT	Patel Teja EOS	

5. References

Absoft (2008). *Absoft FORTRAN-95 compiler 10.1*. Computer program. Rochester Hills, Michigan: Absoft Corporation.

Compaq (2000). *Compaq Visual Fortran language reference*. Houston: Compaq Computer Corporation.

Deiters, U.K. (2002). Calculation of densities from cubic equations of state. *AIChE Journal* 48(4), 882-886.

Deiters, U.K. (2005). Reply to letter to the editor. *AIChE Journal* 51(12), 3310.

Jenkins, M.A. (1975). Algorithm 493: Zeros of a Real Polynomial. *ACM Transactions on Mathematical Software* 1(2), 178-189.

Jenkins, M.A. & Traub, J.F. (1970). A three-stage algorithm for real polynomials using quadratic iteration. *SIAM Journal on Numerical Analysis* 7(4), 545-566.

Jenkins, M.A. & Traub, J.F. (1975). Principles for testing polynomial zero finding programs. *ACM Transactions on Mathematical Software* 1(1), 26-34.

Kim, H., Lin, H.M. & Chao, K.C. (1986). Cubic chain-of-rotators equation of state. *Industrial and Engineering Chemistry Fundamentals* 25(1), 75-84.

Koçak, M.Ç. (2008). A class of iterative methods with third-order convergence to solve nonlinear equations. *Journal of Computational and Applied Mathematics* 218(2), 290-306.

Koçak, M.Ç. (2008a). Simple geometry facilitates iterative solution of a nonlinear equation via a special transformation to accelerate convergence to third order. *Journal of Computational and Applied Mathematics* 218(2), 350-363.

Liley, P.E., Thomson, G.H., Friend, D.G., Daubert, T.E. & Buck, E. (1999). Physical and Chemical Data: Critical constants. In: Perry, R.H., Green, D.W. & Maloney, J.O. (editors). *Perry's Chemical Engineers' Handbook.* 7th ed. New York: McGraw-Hill, Inc., (Chapter 2).

Lin, H.M., Kim, H.Y., Guo, T.M. & Chao, K.C. (1983). Cubic chain-of rotators: equation of state and VLE calculations. *Fluid Phase Equilibria* 13, 143-152.

Mathias, P.M. & Benson, M.S. (1986). Computational aspects of equations of state: fact and fiction. *AIChE Journal* 32(12), 2087-2090.

Mathias, P.M., Boston, J.F. &Watanasiri, S. (1984). Effective utilization of equation of state for thermodynamic properties in process simulation. *AIChE Journal* 30(2), 182-186.

Orbey, H. & Sandler, S.I. (1998). *Modeling vapor-liquid equilibria: Cubic equations of state and their mixing rules*. Cambridge: Cambridge University Press.

Peng, D.Y. & Robinson, D.B. (1976). A new two-constant equation of state. *Industrial & Engineering Chemistry Fundamentals* 15(1), 59-64.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., & Flannery, B.P. (1992). *Numerical Recipes in Fortran. The art of scientific computing*. Cambridge (United Kingdom): Cambridge University Press.

Salim, P.H. (2005). Letter to the editor. *AIChE Journal* 51(12), 3309.

Salim, P.H. (2006). Comment on the paper of Zhi and Lee entitled “Fallibility of analytic roots of cubic equations of state in low temperature region. *Fluid Phase Equilibria* 240(2), 224-226.

Scilab (2009). *Scilab 5.1.*Computer program. Rocquencourt, France: The Scilab Consortium (DIGITEO)

Teja, A. &Patel, N.C. (1982). A new cubic equation of state for fluids and fluid mixtures. *Chemical Engineering Science* 37(3), 463-473.

Valderrama, J.O. (2003). The state of the cubic equations of state. *Industrial and Engineering Chemistry Research* 42(8), 1603-1618.

Valderrama, J.O. & Alfaro, M. (2000). Liquid volumes from generalized cubic equations of state: take it with care. *Oil & Gas Science and Technology* 55(5), 523-531.

van der Waals, J.D. (1873). *Over de continuïteit van den gas- en vloeistoftoestand (On the continuity of the gas and liquid state)*. Doctoral Thesis, Leiden University, Leiden, The Netherlands.

Weisstein, E.W. (2009). *Cubic formula*. http://mathworld.wolfram.com/CubicFormula.html

Weisstein, E.W. (2009a). *Halley's method.* http://mathworld.wolfram.com/HalleysMethod.html

Weisstein, E.W. (2009b). *Vieta's substitution*., http://mathworld.wolfram.com/VietasSubstitution.html

Wong, D.S.H., Orbey, H. & Sandler, S.I. (1992). Equation of state mixing rule for nonideal mixtures using available activity coefficient model parameters and that allows extrapolation over large ranges of temperature and pressure. *Industrial & Engineering Chemistry Research* 31(8), 2033-2039.

ZareNezhad, B. &Eggeman, T. (2006). Application of Peng-Robinson equation of state for $_2CO$ freezing prediction of hydrocarbon mixtures at cryogenic conditions of gas plants.*Cryogenics* 46(12), 840-845.

Zhi, Y. & Lee, H. (2002). Fallibility of analytic roots of cubic equations of state in low temperature region. *Fluid Phase Equilibria* 201(2), 287-294