ELSEVIER

Mathematics and Computers in Simulation 48 (1998) 75–91

MATHEMATICS AND COMPUTERS IN SIMULATION

Identifying and removing sources of imprecision in polynomial regression

Neima Brauner[a,*], Mordechai Shacham[1,b]

[a]*School of Engineering, Tel-Aviv University, Tel Aviv 69987, Israel*
[b]*Department of Chemical Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel*

Abstract

Identification and removal of imprecision in polynomial regression, originating from random errors (noise) in the independent variable data is discussed. The truncation error-to-noise ratio (TNR) is used to discriminate between imprecision dominated by collinearity, or numerical error propagation, or inflated variance due to noise in the independent variable. It is shown that after the source of the imprecision has been identified, it can often be removed by simple data transformations or using numerical algorithms which are less sensitive to error propagation (such as QR decomposition). In other cases, more precise independent variable data may be required to improve the accuracy and the statistical validity of the correlation.

Keywords: Regression; Polynomial; Precision; Noise; Collinearity

1. Introduction

Mathematical modeling and simulation of physical phenomena requires, in addition to an accurate model, precise equations to represent pertinent physico-chemical properties as a function of state variables, such as temperature, pressure, and composition. The accuracy of simulations of physical phenomena critically depends on the accuracy of these correlation equations. Such equations require fitting some parameters by regression of experimental data.

A general form of a regression model is

$$y = F(\mathbf{x}, \boldsymbol{\beta}) + \epsilon \tag{1}$$

where y is the dependent variable, $\mathbf{x}$ represents a vector of independent (state) variables, $\boldsymbol{\beta}$ is a vector of parameters to be fitted to the model and ϵ is (random) error in the measured y values.

* Tel.: +972-3-6408127; fax: +972-3-6407334; e-mail: brauner@eng.tau.ac.il

[1]Tel.: +972-7-6461481; fax: +972-7-6472916; e-mail: shacham@bgumail.bgu.ac.il

0378-4754/98/$ – see front matter
PII: S0378-4754(98)00146-3

Assuming the model is correct, modern regression techniques allow arriving at equations and parameter estimates which can predict values within the experimental error, ϵ. However, several causes may prevent reaching this goal. These are related to the noise in the independent variables data due to the unavoidable limited precision in their reported values. This noise is eventually transformed to imprecision of the calculated parameter values due to numerical error propagation and collinearity amongst the (presumably) independent variables.

Unfortunately, the effects of collinearity and numerical error propagation have not been taken into account in published correlations of various thermophysical properties (see for example [1] or [2]). As a result, the correlations may either contain an insufficient number of parameters to represent the data accurately or too many parameters. If there are too many parameters, the correlation becomes ill-conditioned, whereby adding or removing experimental points from the data set may drastically change the parameter values. Also, derivatives are not represented correctly and extrapolation may yield absurd results even for a small range of extrapolation. Collinearity in polynomial regression was addressed, for example, by Bradley and Srivastava [3] and Seber [4], who discuss the problems that can be caused by collinearity and suggested certain measures that can be taken in order to reduce the undesired effects of collinearity. Shacham and Brauner [5] have proposed an indicator to diagnose collinearity and suggested several data transformations to reduce collinearity and its undesired effects in polynomial regression. The use of this collinearity diagnostic has been recently extended to other types of regression [6,7].

In order to increase the precision of a correlation it is necessary first to identify the dominant cause for the imprecision. Then, appropriate measures for alleviating its effects and increasing the correlation precision can be applied.

In this paper, indicators for discriminating among the possible causes of imprecision, which are related to independent variable data, are presented and methods for alleviating their effects are suggested. The proposed methods are demonstrated in examples involving regression of vapor pressure and heat capacity data. The discussion is limited to polynomial regression, but the results can be readily extended to other forms of regression models.

The calculations were carried out using the student edition of MATLAB [8] and POLYMATH 4.0 [9] packages.

2. Problem definition

Let us assume that there is a set of N data points of a dependent variable y_i versus an independent variable x_i. A nth order polynomial fitted to the data is of the form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_n x_i^n + \epsilon_i \tag{2}$$

where β_0, β_1,...,β_n are the parameters of the model and ϵ_i is the error in y_i. The vector of estimated parameters $\hat{\boldsymbol{\beta}}^{\mathrm{T}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n)$ is often calculated via the least squares error approach, by solving the normal equation:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{3}$$

The rows of $\mathbf{X}$ are $\mathbf{x}_i = 1,\ x_i, \ldots, x_i^n$ and $\mathbf{X}^{\mathrm{T}}\mathrm{X}{=}\mathbf{A}$ is the normal matrix. Another method to obtain the

parameter values in Eq. (2) is by solving the following over determined system of equations:

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} \tag{4}$$

using QR decomposition [10]. The QR decomposition method requires more arithmetic operations than the solution of the normal equations but it is less sensitive to numerical error propagation.

A numerical indicator for the quality of the fit which is used most frequently is the square of standard error of the estimate, which represents the sample variance, and is given by

$$s^2 = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N - n - 1} \tag{5}$$

Thus, the sample variance is the sum of squares of errors divided by the degrees of freedom (where the number of parameters, $n+1$, is subtracted from the number of data points, N) and is a measure for the variability of the actual $\hat{y}$ values. Smaller variance indicates a better fit of the model to the data.

One of the assumptions of the least squares error approach is that there is no error in the independent variables. However, this is rarely true. The precision of independent variables is limited due to limitations of the measuring and control devices. Thus, the value of an independent variable can be represented by

$$x_i = \tilde{x}_i + \delta_i \tag{6}$$

where $\tilde{x}_i$ is the expected value of the measured x_i and δ_i is the error (uncertainty, noise) in its value. The least squares error approach can be applied in a way that considers the error in both the dependent and independent variables (see, for example, p. 87 in [11]), but this will usually have a very little effect on the calculated values of $\hat{\boldsymbol{\beta}}$. Nevertheless, the error in the independent variable plays an important role in determining the highest degree of the polynomial and the number of parameters that can be fitted to the data. The highest degree of a polynomial that can be used for a particular set of data is often related to collinearity among the terms of Eq. (2).

3. Collinearity and its diagnostics

Collinearity plays an important role in limiting the precision of a correlation. In polynomial regression, collinearity is said to exist among the columns of $\mathbf{X} = [1, \mathbf{x}, \mathbf{x}^2, \ldots, \mathbf{x}^n]$, if for a suitable small predetermined $\eta>0$, there exist constants c_0, c_1, c_2, …, c_n not all of which are zero, such that

$$c_0 + c_1\mathbf{x} + c_2\mathbf{x}^2 + \cdots + c_n\mathbf{x}^n = \boldsymbol{\Delta}, \quad \text{with } \|\boldsymbol{\Delta}\| < \eta\|c\| \tag{7}$$

This definition cannot be used directly for diagnosing collinearity because it is not known how small η should be so that the harmful effects of collinearity will show. Belsley [12] lists several criteria and procedures that can be used to detect collinearity. From among those in this list, the following indicators will be briefly reviewed: the condition number of the normal matrix ($\kappa(\mathbf{A})$), variance inflation factor (VIF) and confidence intervals. The collinearity indicator, truncation error-to-noise ratio (TNR), recently introduced by Shacham and Brauner [5], will also be described.

3.1. Condition number of the normal matrix, $\kappa(\mathbf{A})$

The condition number, $\kappa(\mathbf{A})$ is often used to estimate the errors introduced into the parameter values due to numerical error propagation. It can be shown that the errors in the calculated parameter values, $\delta\hat{\boldsymbol{\beta}}$ are bounded by (p. 176 in [13]):

$$\kappa(\mathbf{A})\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|} \geq \frac{\|\delta\hat{\boldsymbol{\beta}}\|}{\|\hat{\boldsymbol{\beta}} + \delta\hat{\boldsymbol{\beta}}\|} \tag{8}$$

where $\delta\mathbf{A}$ is the matrix of errors in $\mathbf{A}$, and $\|\cdot\|$ indicates the norm of a matrix or a vector. A similar equation relates the error in $\mathbf{b}$, $\delta\mathbf{b}$, to the error $\delta\hat{\boldsymbol{\beta}}$:

$$\frac{\|\delta\hat{\boldsymbol{\beta}}\|}{\|\hat{\boldsymbol{\beta}}\|} \leq \kappa(\mathbf{A})\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \tag{9}$$

The condition number is the ratio of the largest to the smallest eigenvalue of $\mathbf{A}$. A strong collinearity results in a higher condition number, thereby amplifying both $\delta\mathbf{b}$ and $\delta\mathbf{A}$. The former represents measurement errors in both the dependent and independent variables while $\delta\mathbf{A}$ represents the errors in the independent variables.

3.2. Variance inflation factor (VIF)

The variance inflation factor can be defined as (p. 27 in [12]):

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2} \tag{10}$$

where R_j^2 is the multiple correlation coefficient of $\mathbf{x}^j$ regressed on the remaining columns of the $\boldsymbol{X}$ matrix. For non-centered data (the β_0 parameter does not vanish) the following equation is used for calculating the multiple correlation coefficient:

$$R_j^2 = 1 - \frac{\sum_i \left(\hat{x}_i^j - x_i^j\right)^2}{\sum_i \left(x_i^j\right)^2} \tag{11}$$

where $\hat{x}_i^j$ is the calculated value of x_i^j when it is regressed on the remaining powers of x_i. A high level of collinearity leads the R_j value close to one, which causes VIF_j to attain a large positive value. The value of VIF_j is calculated for j=0, 1, …, n (all the columns of matrix $\mathbf{X}$). The maximal VIF_j is usually used for collinearity diagnostic. The term VIF will be used for the maximal VIF_j henceforth.

Both $\kappa(\mathbf{A})$ and VIF rely only on independent variable data for diagnosing collinearity. The disadvantage of these indicators is that there are no well established threshold values for them to indicate harmful level of collinearity. Therefore, they cannot be considered as quantitative measures for collinearity.

3.3. Confidence intervals

A frequently used statistical indicator to determine whether a particular term should be included in the model is the confidence interval. This interval is defined by

$$\hat{\beta}_j - t(\nu,\alpha)s\sqrt{a_{jj}} \le \beta_j < \hat{\beta}_j + t(\nu,\alpha)s\sqrt{a_{jj}}, \tag{12}$$

where $t(\nu,\alpha)$ is the statistical t distribution corresponding to ν degrees of freedom and a desired confidence level, α and s is the standard error of the estimate.

The confidence interval test relies on more information than is required for the previous tests. In particular, it depends on s, which reflects the measurement errors in the dependent variable (and also, indirectly, the error in the independent variable) and the magnitude of the diagonal elements of $\mathbf{A}^{-1}$, which strongly relate to $\kappa(\mathbf{A})$.

Clearly, if $\hat{\beta}_j$ is smaller in its absolute value than the term $t(\nu,\alpha)s\sqrt{a_{jj}}$, then the zero value is included inside the confidence interval, $\pm t(\nu,\alpha)s\sqrt{a_{jj}}$. Thus, there is no statistical justification to include the associated term in the regression model. If the independent variables are strongly correlated, most confidence intervals will be larger (in absolute values) than the respective parameter values. Thus, the confidence interval test may be insufficient to pinpoint which of the terms should be removed from the model due to collinearity.

Confidence intervals are useful for evaluating the statistical significance of a regression model. However, the calculation of the confidence intervals requires carrying out first the experiments (for obtaining the values of the independent variables) and then the regression calculations. Also, since the values of the confidence intervals depend on several factors, they are not useful in identifying the dominant cause that limits the precision of the regression model.

3.4. Truncation error-to-noise ratio (TNR)

Let us consider Eq. (7) which was used to define collinearity. This equation can be divided by, say, c_j to yield:

$$c_{0,j} + c_{1,j}\mathbf{x} + c_{2,j}\mathbf{x}^2 + \cdots + \mathbf{x}^j + \cdots + c_{n,j}\mathbf{x}^n = \mathbf{\Delta}_j, \tag{13}$$

where $c_{k,j} = c_k/c_j,\ k = 1,2,\ldots,n$. When the coefficients c are obtained by regressing $\mathbf{x}^j$ as a function of the other independent variables, $\mathbf{\Delta}_j$ is the residual of this representation and is denoted as the ''truncation error''. Since the independent variables are subject to an error, the value of $\mathbf{\Delta}_j$ is also subject to an error.

In calculating the error in $\mathbf{\Delta}_j$ *(denoted* $\boldsymbol{\delta}_j$*)*, two cases are considered. When the errors in the various powers of $\mathbf{x}$ are uncorrelated, as is for a noise caused by a limited numerical precision of the computer, then the general error propagation formula (applied to Eq. (13)) yields $\delta_{i,j} = \sum_{k=1}^{n} k|c_{k,j}||x_i^{k-1}\delta_i|$. On the other hand, when the errors are correlated (the measurement noise in x_i is carried out to its various powers), the expression obtained for $\boldsymbol{\delta}_j$ is $\delta_{i,j} = \sum_{k=1}^{n} c_{k,j}kx_i^{k-1}|\boldsymbol{\delta}_i| \le |jx_i^{j-1}||\boldsymbol{\delta}_i|$. Thus, the r.h.s. of this inequality is considered as an error estimate for this case. Consequently, the TNR in polynomial regression can be defined as [5]:

$$\mathrm{TNR_j} = \frac{\|\Delta_j\|}{\|\boldsymbol{\delta j}\|} \tag{14}$$

The value of TNR_j expresses how much of the variation of the truncation error about the zero (mean) value (as represented by the residual plot) can be attributed to experimental 'noise'. When $\mathrm{TNR}_j \le 1$, the truncation error is not significantly different from zero, thus there is a harmful collinearity among the

regressors used in the model. Whereas a value of $TNR_j \gg 1$, indicates that the truncation error is much larger than the noise level, thus harmful effects of collinearity are not expected. In between those two extremes, numerical experimentation can determine the critical values of TNR_j.

A noise in the dependent variable data increases the variance. In order to determine the contribution of the noise in $\mathbf{x}$ to the error in the calculated values of y (thus to the variance), truncation error-to-noise ratio for y, TNR_y can similarly be derived. To calculate TNR_y, the model parameters are calculated after a randomly distributed error of a magnitude $\|\delta\|$ is introduced into the independent variable. The resulting parameter values are denoted by $\boldsymbol{\beta}(x+\delta)$ and the corresponding predicted y values, $\hat{\mathbf{y}}(x+\delta)$. The TNR_y is defined as

$$\mathrm{TNR_y} = \frac{\|(\hat{\mathbf{y}}_{(x)} - \mathbf{y})^2\|}{\|(\hat{\mathbf{y}}_{(x+\delta)} - \hat{\mathbf{y}}_{(x)})^2\|} = \frac{\|\mathbf{s}^2_{(x)}\|}{\|\mathbf{s}^2_{(x+\delta)} - \mathbf{s}^2_{(x)}\|} \tag{15}$$

As indicated by Eq. (15), TNR_y represents the ratio between the total variance to that part of the variance which is due to the noise in $\mathbf{x}$. A value of $TNR_y \gg 1$ indicates that the contribution of the noise, the error in y is not very significant and the correlation variance is dominated by the noise in y or by the lack of fit of the model. On the other hand, a TNR_y value close to one, or smaller than one, indicates that the noise in x dominates and has the major contribution to the imprecision of the model.

4. Transformations to reduce collinearity

These transformations have been analyzed in detail by Shacham and Brauner [5]. Only the main results of this analysis are described here. There are several transformations that can be used to reduce collinearity. A transformation which is routinely used is division of the values of x_i by $x_{\max}$, where $x_{\max}$ is the point with the largest absolute value. Thus, $v_i = x_i/x_{\max}$, where v_i is the normalized x_i value. If all the x_i are of the same sign (say $x_i > 0$), then the normalized value will vary in the range $0 < v_{\min} \leq v_i \leq 1$. The v-transformation can reduce considerably the value of the condition number. It has no effect, however, on the level of the collinearity and it will not change the value of the indicators VIF, TNR and TNR_y.

The following w- and z-transformations can significantly reduce the level of collinearity in polynomial regression:

The w-transformation $w_i = (x_i - x_{\min})/(x_{\max} - x_{\min})$ yields variable distribution in the range $0 \leq w_i \leq 1$. This type of transformation was used by Wagner [14], for example, to develop most accurate correlations for vapor pressure.

The z-transformation $z_i = (2x_i - x_{\max} - x_{\min})/(x_{\max} - x_{\min})$ yields variable distribution in the range of $-1 \leq z_i \leq 1$. Similar transformations are widely used and highly recommended by statisticians (see, for example, [4]).

In the following sections, three examples will be presented where the various collinearity indicators and TNR_y are used to identify the sources of the imprecision in correlations. Data transformations and solution of the least squares equation by QR decomposition are used to remove the dominant sources of the imprecision.

5. Example 1: Effects of numerical error propagation

One of the widely known effects of collinearity is that it can intensify numerical error propagation up to a harmful level. In this example, the consequences of a harmful level of numerical error propagation in polynomial regression are demonstrated.

In order to enable discrimination between the various possible sources of inaccuracy, "exact" dependent variable data will be generated using a model equation. In this way, a random experimental noise in $\boldsymbol{y}$ is avoided. Vapor pressure data of toluene is generated by the Wagner equation [14]:

$$\ln P_R = \frac{1}{T_R}[\beta_0(1-T_R) + \beta_1(1-T_R)^{1.5} + \beta_2(1-T_R)^3 + \beta_3(1-T_R)^6] \tag{16}$$

where $T_R = T/T_c$ is the reduced temperature, $P_R = P/P_c$ the reduced pressure, T the temperature (K), P the pressure (kPa), T_c the critical temperature (K) and P_c is the critical pressure (kPa).

Table 1 shows the critical constants and the Wagner equation coefficients for toluene. Using Eq. (16) with the parameters from Table 1, 41 equally spaced data points were generated in a 100 K temperature range. Minimal and maximal values of $Y\,(=\ln P/\ln(P_{max}))$ and $v\,(=T/(T_{max}))$ are shown in Table 2. Polynomials of the form of Eq. (2) were fitted to the data using the v-, w- and z-transformations.

The solution for the coefficients of polynomials were obtained using both matrix inversion (Eq. (3)) and QR decomposition (Eq. (4)). It should be noted that, because of the $(1-T_R)^{1.5}$ term, the Wagner equation cannot be represented exactly even by a very high order of polynomial with integer exponents. However, increasing the order of such a polynomial should make the representation more accurate, resulting in smaller variance. Fig. 1 shows the variance as a function of the polynomial order for the v-, w- and z-transformations when matrix inversion is used and the variance when the z-transformation is used with QR decomposition. With the v-transformation, a steady decrease of the variance up to the fourth order polynomial is obtained, whereby each additional term in the polynomial reduces the variance by about two orders of magnitude.

Starting at the fifth order polynomial, the normal matrix becomes ill-conditioned, resulting in a sharp increase of the variance. With the w- and z-transformations, however, the same rate of decrease of the variance extends to the sixth and the ninth order polynomials, respectively. For the w-transformation, the variance increase starts at the seventh order polynomial, while for the z-transformation, the variance starts to increase at the ninth order polynomial. Using the QR decomposition (with the z-trans-

Table 1
Critical properties and the Wagner equation constants for toluene [15]

Melting point temperature[a] (K)	178.15
Normal boiling point temperature[a] (K)	383.75
Critical temperature (K)	591.72
Critical pressure (kPa)	4106.45
Wagner constants	
β_0	−7.28607
β_1	1.38091
β_2	−2.83433
β_3	−2.79168

[a]Ref. [16].

Table 2
Minimal and maximal values for the first vapor pressure data set

Variable	Value
T_{max} (K)	433.75
T_{min} (K)	333.75
v_{max}	1.0
v_{min}	0.769452
P_{max} (kPa)	347.812
P_{min} (kPa)	18.9721
Y_{max}	1.0
Y_{min}	0.50293

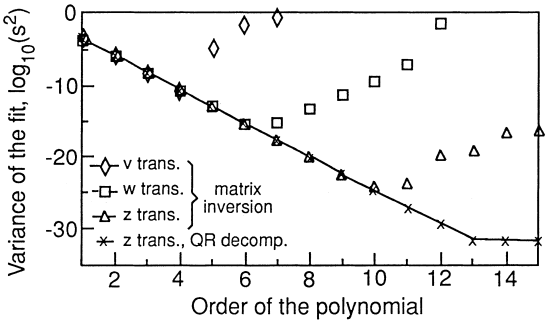

Fig. 1. Variances of various order polynomials using v-, w- and z-transformations and QR decomposition.

formation) the minimum of the variance is reached at the 13th order polynomial. Thus minimizing the collinearity by using the z-transformation and minimizing the numerical error propagation using QR decomposition enables reducing the variance of the best fit to 2.81×10^{-32} using 13th order polynomial from the value of 4.41×10^{-11} obtained with the fourth order polynomial, using v-transformation and matrix inversion.

The ill effects of collinearity related to numerical error propagation can be further demonstrated with reference to Table 3. In this table, the results obtained for the fifth and sixth order polynomials, using v-transformation are summarized. The parameter values, standard errors and variances obtained with QR decomposition and inversion of the normal matrix are compared.

The results obtained with the QR decomposition are much more accurate than those obtained by matrix inversion. For the fifth order polynomial the QR decomposition yields a variance of 1.3×10^{-13}, while with matrix inversion the variance is 1.69×10^{-5}. Increase of the variance by eight orders of magnitude causes an increase of four orders of magnitude in the standard errors (s.e.) of β, which become much larger than the parameter values themselves.

Fig. 2 shows the effect of numerical error propagation due to collinearity on the residuals, obtained with the fifth order polynomial. When QR decomposition is used, the residuals are of the order of 10^{-7} and exhibit an oscillatory pattern around zero, indicating that the use of a higher order polynomial will improve the fit and reduce the variance. With matrix inversion the residuals are of the order of 10^{-3}, all values are negative and their pattern indicates a systematic error in the model.

Table 3
Parameter values, standard errors and variances for the fifth and sixth order polynomials using v-transformation

Index	Fifth order polynomial				Sixth order polynomial	
	QR decomposition		Matrix inversion		QR decomposition	Matrix inversion
	β[a]	s.e. β	β	s.e. β	β	β
0	−11.77	0.034568	−11.769	395.16	−14.824	−10.081
1	46.138	0.19676	46.135	2249.2	67.02	34.694
2	−74.026	0.44717	−74.018	5111.8	−133.42	−41.339
3	64.812	0.50722	64.807	5798.2	154.75	15.449
4	−29.841	0.28714	−29.835	3282.4	−106.33	12.111
5	5.6865	0.064905	5.6859	741.95	40.326	−13.224
6					−6.5255	3.5887
Variance	1.30E−13		1.69E−05		5.95E−16	0.025313

[a] Numbers rounded to five significant digits.

In spite of the dramatic differences in the variances, in the standard errors on the parameters and in the residual plots, the difference in the calculated parameter values is relatively small. The values calculated by the two methods differ only in the fourth decimal digit.

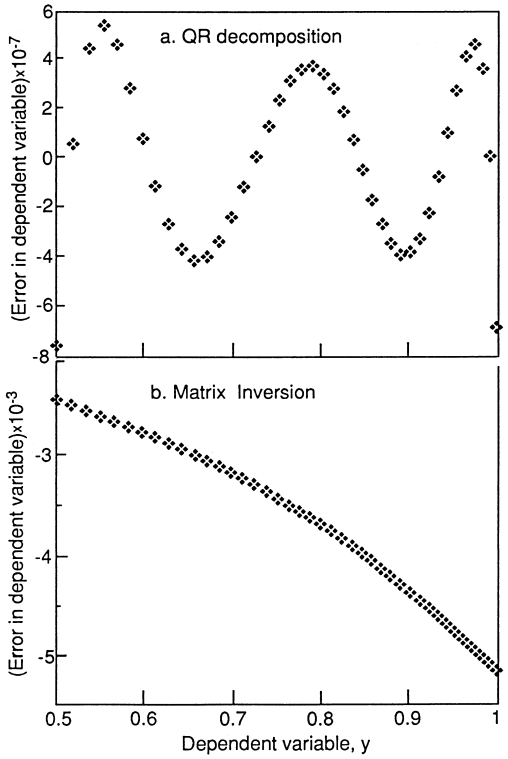

Fig. 2. Residual plots for fifth order polynomial representation of the vapor pressure data using v-transformation.

For the sixth order polynomial, the variance is reduced to 5.95×10^{-16} when QR decomposition is used and increased to 0.0253 with matrix inversion. Standard error of β cannot be calculated because negative values are obtained in the inverted normal matrix and there is not even a single accurate digit in the parameter values which are calculated using the inverse of the normal matrix. Thus the results obtained for the sixth order polynomial provide a very dramatic demonstration of the ill effects of collinearity.

Are the various collinearity indicators capable of predicting when collinearity related to numerical error propagation has reached a harmful level? To answer this question, the various collinearity indicators were plotted versus polynomial order. Fig. 2 shows $\log(\kappa(\mathbf{A}))$ as a function of the polynomial order for various transformations. The logarithm of the condition number increases linearly as the polynomial order increases (linearity is distorted when numerical error propagation prevents obtaining accurate values for $\kappa(\mathbf{A})$, as is the case for the sixth order polynomial with v-transformation). The rate of increase of $\log(\kappa(\mathbf{A}))$ with the polynomial order is, as predicted by Shacham and Brauner [5] on the theoretical basis, approximately 3 for the v-transformation, 1.5 for the w-transformation and 0.72 for the z-transformation.

Comparing Fig. 3 with Fig. 1 reveals that harmful effects of collinearity show up for different values of $\kappa(\mathbf{A})$ using different data transformations. For the v-transformation, the increase of the variance starts at the fourth order polynomial, where $\kappa(\mathbf{A})=6.71\times10^{11}$. For the w-transformation, it starts increasing at the sixth order polynomial, where $\kappa(\mathbf{A})=3.77\times10^{8}$ and for the z-transformation the variance increases at the 10th order polynomial, where $\kappa(\mathbf{A})=8.33\times10^{6}$. Thus, the condition number alone cannot be used to predict the polynomial order for which numerical error propagation due to collinearity reaches a harmful level that causes the precision of the model to deteriorate.

Table 4 shows the maximal VIF values for the same order of polynomials. It can be seen that the maximal VIF shows exactly the same trend as $\kappa(\mathbf{A})$, thus, it also cannot precisely diagnose harmful effects of collinearity.

Fig. 4 shows the TNR values for various order of polynomials using the three transformations. The TNR values were calculated using an uncorrelated noise level of $\delta T=5\times10^{-5}$ K in the temperature data. This level of noise was found by numerical experimentation to represent the effective error level in the matrix inversion algorithm (for the QR decomposition algorithm the effective error is much smaller).

It can be seen that the various curves cross the line of TNR=1 (log (TNR)=0) at the following points: fourth order polynomial for the v-transformation, sixth order polynomial for the w-transformation, and

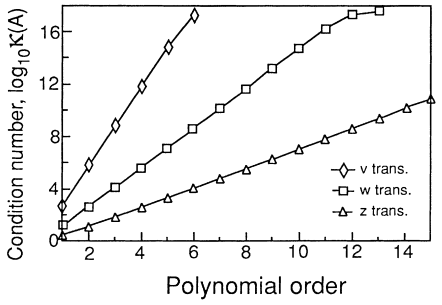

Fig. 3. Condition number as a function of polynomial order for v-, w- and z-transformations.

Table 4
Condition number, maximal VIF and TNR values for polynomial orders where the variance starts to increase

Transformation	Order of the polynomial	$\kappa(\mathbf{A})$	Maximum VIF	TNR
v	4	6.71×10^{11}	6.44×10^{10}	0.883
w	6	3.77×10^{8}	1.26×10^{8}	0.621
z	10	8.33×10^{6}	2.7×10^{5}	3.11

between 10th and 11th order polynomial for the z-transformation. Thus, the TNR predicts correctly the order of the polynomial at which increase of the variance (due to numerical error propagation) starts. This conclusion is further reinforced by comparing the TNR, maximum VIF and $\kappa(\mathbf{A})$ values in Table 4.

6. Example 2: Effects of range reduction

To investigate the effects of random noise in the independent variable data, ''exact'' vapor pressure data was generated, using the following fourth order polynomial (obtained by regressing the data in Table 2):

$$\ln P = \beta_0 + \beta_1 T + \beta_2 T^2 + \beta_3 T^3 + \beta_4 T^4 \tag{17}$$

with the parameters β_0=−51.190672, β_1=0.3302147, β_2=−0.0010848467; β_3=1.4620472×10^{-6} and β_4=−7.74648805×10^{-10}. With this polynomial, ''exact'' vapor pressure data was generated in five different temperature ranges, t-range=100, 80, 60, 40 and 20 K, each range includes 41 equally spaced data points. After generating the ''exact'' vapor pressure data, normally distributed random noise of a magnitude of δT=0.005 K was introduced to the temperature data. Polynomials of second, third and fourth order were fitted to the vapor pressure data, using the v-transformation for the temperature data and normalized vapor pressure data $(Y = \ln(P)/\ln(P_{\max}))$. The change of the variance of the fit as a function of the temperature range is shown in Fig. 5.

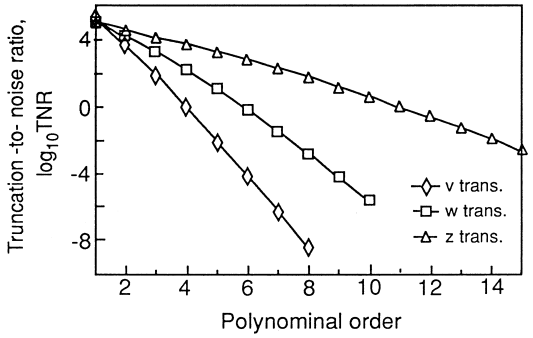

Fig. 4. TNR values for various order polynomials for noise level of δT=5×10^{-5} K.

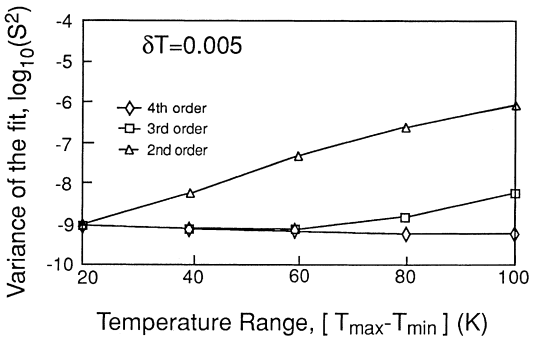

Fig. 5. Change of variance as a function of temperature range for δT=0.005 K.

The variance for the second order polynomial demonstrates a typical behavior, where a lack of fit caused by an inappropriate model (low polynomial order) dominates over the lack of fit affected by the error introduced in the temperature data. In this case, reducing the range affects a sharp decrease of the variance value. This is expected since data that were generated by a fourth order polynomial can be represented by a second order polynomial much better in a narrower range of 20 K than in a wider range of 100 K.

The variance for the fourth order polynomial demonstrates a behavior where the dominant cause for imprecision is the noise in the temperature values. A reduction of the range does not reduce the value of the variance, but on the contrary, it slightly increases. For the 20 K range, the variance is the same for the second, third and fourth order polynomials, and up to a range of 60 K the variance is almost unchanged for the third and fourth order polynomials.

Additional aspects of the difference between lack of fit of the model (caused by insufficient number of terms in the polynomial) and that caused by noise in the independent variable data are demonstrated in Table 5. In this table, parameter values and variances of the second and fourth order polynomials are shown for noise levels of δT=0 (no noise), δT=0.001 K and δT=0.01 K. It can be seen that when a second order polynomial is used, the introduction of a noise, has a modest effect. With δT=0.001 K, there are still four accurate digits remaining in the parameter values, and three accurate digits for δT=0.01 K. The variance with δT=0.01 K increases by a factor of about 40 compared to variance obtained with δT=0. In contrast to these modest changes, the fourth order polynomial exhibits severe

Table 5
Parameter values and variances of second and fourth order polynomials with various temperature noise levels

	Second order polynomial			Fourth order polynomial		
	δT=0[a]	δT=0.001	δT=0.01	δT=0	δT=0.001	δT=0.01
β_0	−3.8774	−3.877	−3.8728	−10.457	−12.653	−35.153
β_1	7.6074	7.6065	7.5985	31.385	40.482	132.81
β_2	−2.7299	−2.7295	−2.7257	−34.357	−48.358	−190.55
β_3				18.232	27.808	125.11
β_4				−38.035	−62.594	−31.228
Vaiance	8.8194E−11	1.2015E−10	3.2664E−09	7.8486E−29	3.3551E−11	3.3548E−09

[a] Numbers rounded to five significant digits.

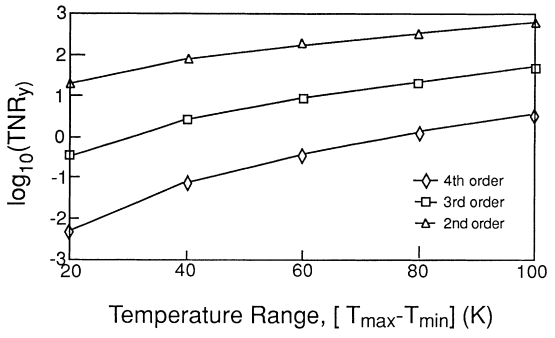

Fig. 6. Change of TNR as a function temperature range for δT=0.005 K.

effects of ill-conditioning when a noise is introduced to the temperature data. It can be seen that for a noise level of δT=0.001 K, there is not even a single correct digit in the parameter values and for δT=0.01 K, many of them change by almost an order of magnitude. With δT=0.01 K, the variance increases by a factor of 0.42×10^{20} compared to its value for δT=0. For such a noise level, the variance of the second order polynomial is lower than the variance of the fourth order polynomial.

The TNR_y indicator (Eq. (15)) can be used to discriminate between the case when the noise in the independent variable limits the precision and the cases where the model is not appropriate (due to insufficient number of terms in the polynomial). In Fig. 6, the TNR_y values for noise level of δT=0.005 is plotted. It can be seen that for the fourth order polynomial, $TNR_y \simeq 3.0$ for the widest range and the lower noise level, but $TNR_y<1$ for most of the other cases with narrower range and/or higher noise level. Thus, TNR_y clearly indicates that the noise in the independent variable is the dominant factor in limiting the precision of the correlation obtained with a fourth order polynomial.

The TNR_y value for the second order polynomial ranges from $\sim$510 for the widest range to about 20 for the narrowest range. Thus, for most of the region the noise level is not the main source of imprecision. In this case, the correlation precision is dominated by a lack of fit of the model (insufficient number of terms in the polynomial). Comparing Figs. 5 and 6 shows that TNR_y is a good measure for the extent to which noise in the independent variable affects the imprecision of the correlation.

The above results indicate that a narrower range requires higher precision of the independent variable, to avoid deterioration of the correlation precision. To keep TNR_y at a constant value when the range is reduced, the noise level must be reduced in proportion to the nth order of the range reduction. In Fig. 7, change of the variance of the various polynomial fits as function of the temperature range are shown. The noise introduced is: δT=0.005 (t-range/100)^n. Comparing Fig. 7 with Fig. 5 shows that when such a noise reduction is associated with range reduction, the noise in the independent variable does not become a limiting factor.

7. Example 3: Correlation of heat capacity data of solid propylene

Heat capacity versus temperature data are usually correlated by polynomials. Daubert and Danner [1], for example, used a fourth order polynomial to correlate heat capacity (C_P) data versus temperature for solid Propylene, as published by Timmermans [17]. This data is shown in Table 6. Suggested

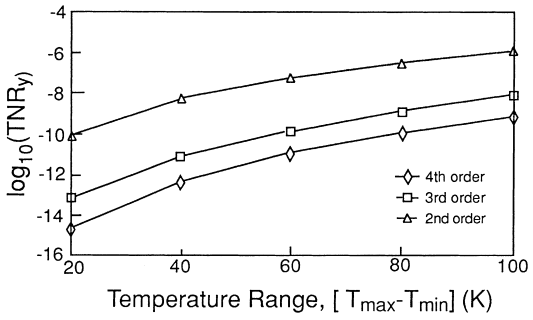

Fig. 7. Change of the variance as a function of temperature range with varying δT=0.005 (t-range/100)^n (n is the polynomial order).

Table 6
Heat capacity data for solid propylene [17]

No.	T (°C)	C_P (cal/(g K))
1	−258.98	0.0271
2	−256.5	0.0378
3	−253.53	0.0526
4	−250.35	0.0725
5	−246.93	0.0926
6	−243.29	0.112
7	−239.54	0.133
8	−235.7	0.153
9	−231.87	0.171
10	−227.63	0.187
11	−222.5	0.207
12	−217.2	0.224
13	−211.91	0.244
14	−206.55	0.261
15	−201.71	0.276
16	−196.67	0.292
17	−196.57	0.295
18	−191.47	0.319
19	−188.29	0.339

validity range of the correlation is 13.0 K$\leq T\leq$87.0 K. The reported precision of the temperature measurements is $\delta T=\pm 0.05$°C. Only the first 19 data points out of 20 provided by Timmermans were used, since the last data point (nearly at the melting point) turned out to be very inaccurate due to premelting.

To verify the precision and the statistical validity of the fourth order polynomial correlation for the heat capacity data, polynomials of up to fourth order were fit to the normalized values of $C_P(\bar{C}_P = C_{P_i}/0.339)$ versus normalized T (v_i=T_i/84.86, the temperature was first converted to K). Fig. 8 shows the residual plots for third and fourth order polynomials. It can be seen that third order polynomial is insufficient for representing this data. There is a clear trend in the residual plot and the

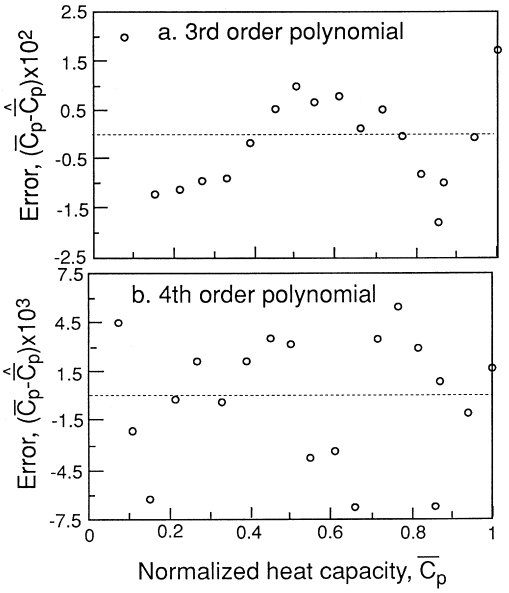

Fig. 8. Residual plots of polynomial representation of the propylene heat capacity data.

maximal error is 25% (for the first data point). The representation by the fourth order polynomial is much better. The residual plot shows randomly distributed errors, with a maximal relative error of 5%.

Table 7 shows the parameter values, the confidence intervals, variances and TNRs for third and fourth order polynomials. The variance decreases considerably from the third to the fourth order polynomial. However, the confidence intervals on the parameter values increase. For the fourth order polynomial, the value of β_1 is no longer significantly different from zero. Consequently, only the use of third order polynomial can be justified on a statistical ground, since for the fourth order polynomial a small change in the data can introduce large changes in the parameter values. Indeed, removing the first data point from the set, yields the following fourth order polynomial parameters: β_0=−0.117042, β_1=0.697793, β_2=3.37451, β_3=−5.9857 and β_4=3.02791. Most of these values differ already in the first significant digit from the values shown in Table 7.

Table 7
Parameters, confidence intervals, variances and TNR values for third and fourth order polynomials representing $\bar{C}_P(v)$

Parameter	Third order polynomial	Fourth order polynomial
β_0	−0.28718±0.06204	−0.075352±0.05447
β_1	2.40499±0.405	0.371259±0.4946
β_2	−2.12093±0.770231	4.25054±1.499
β_3	0.984937±0.4411	−6.95602±1.839
β_4		3.40788±0.7857
Variance	0.0001285	1.92E−05
TNR	14.01	2.608
TNR_y	26.94	16.99

Table 8
Parameters, confidence intervals, variances and TNR values for third and fourth order polynomials representing $\bar{C}_P(z)$

Parameter	Third order polynomial	Fourth order polynomial
β_0	0.589678±0.009	0.601582±0.004447
β_1	0.389812±0.02347	0.38934±0.009126
β_2	−0.068842±0.01669	−0.167284±0.0236
β_3	0.0711677±0.03188	0.0720998±0.01239
β_4		0.102561±0.02365
Variance	0.0001285	1.92E−05
TNR	76.07	33.62
TNR_y	26.94	16.99

The TNR value for the fourth order polynomial (shown in Table 7) is 2.6, while TNR_y=16.99. A value of TNR close to 1 (2.6), indicates that collinearity among the various powers of the independent variables is a possible cause for the wide confidence intervals. If this is the case, the level of collinearity can be reduced using the z-transformation. Table 8 shows the parameter values, the confidence intervals, variances and TNRs for third and fourth order polynomials, representing $\bar{C}_P$ as a function of the transformed variable z. It can be seen that for both the third and fourth order polynomials, all the parameters are significantly different from zero, making both correlations statistically valid. Thus, the z[-1,1] transformation eliminates in this particular case, the undesired effects of collinearity and TNR enables identification of this source of imprecision.

8. Summary and conclusions

Theoretical analysis and numerical experimentation have been used to investigate the various effects of random error (noise) in the independent variable data in polynomial regression.

It has been demonstrated that this noise may often lead to an ''ill-conditioned'' problem, where a small change in the data affects large changes in the parameter values and increasing the order of the polynomial yields a higher variance. Such ''ill-conditioned'' systems result from collinearity and numerical error propagation. A noise in the independent variable data may also cause inflated confidence intervals rendering a particular model statistically invalid.

Two new indicators were used to diagnose the source of the imprecision in the correlation. The TNR, which has been introduced by Brauner and Shacham [5], is used to measure collinearity among the various powers of the independent variable. This indicator can identify collinearity as a source of excessive numerical error propagation and inflated confidence intervals. The TNR_y indicator introduced in this paper, can help in identifying cases where no harmful level of collinearity exists, but the noise in the independent variable causes inflated variance and limits the order of the polynomial that can be used for correlating the dependent variable data.

It has been shown that when the imprecision is caused by numerical error propagation, the use of data transformations (such as w- or z-transformations) and the use of QR decomposition, instead of matrix inversion, can alleviate and even eliminate the problems. Inflated confidence intervals, caused by collinearity (due to limited precision in the reported values of the independent variable) can be reduced considerably by using the w- or z-transformations. However, inflated variances caused by the noise in the independent variable as indicated by $TNR_y<1$ cannot be reduced by numerical manipulations. Their

reduction requires additional measurements of higher precision and/or increasing the range of the measurements.

References

[1] T.E. Daubert, R.P. Danner, Physical and Thermodynamic Properties of Pure Chemicals: Data Compilation, Hemisphere, New York, 1989.
[2] R.C. Reid, J.M. Prausnitz, T.K. Sherwood, Properties of Gases and Liquids, 3rd ed., McGraw-Hill, New York, 1977.
[3] R.A. Bradley, S.S. Srivastava, Correlation in polynomial regression, Amer. Stat. 33 (1979) 11–14.
[4] G.A.F. Seber, Linear Regression Analysis, Wiley, New York, 1977.
[5] M. Shacham, N. Brauner, Minimizing the effects of collinearity in polynomial regression, Ind. Eng. Chem. Res. 36(10) (1997) 4405–4412.
[6] N. Brauner, M. Shacham, Role of range and precision of the independent variable in regression of data, AICHE J. 44(3) (1998) 603–611.
[7] M. Shacham, N. Braunder, A collinearity diagnostic based on truncation error to noise ratio (1998), submitted for publication.
[8] Math Works, Inc., MATLAB, (The Student Edition) Prentice Hall, Englewood Cliffs, NJ, 1992.
[9] M. Shacham, M.B. Cutlip, POLYMATH 4.0 User's Manual, CACHE Corporation, Austin, TX, 1996.
[10] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flanerry, Numerical Recipes in FORTRAN, The Art of Scientific Computing, 2nd ed., Cambridge University Press, Cambridge, 1992.
[11] J. Mandel, Evaluation and Control of Measurements, Quality and Reliability, Marcel Dekker, New York, 1991.
[12] D.A. Belsley, Condition Diagnostics, Collinearity and Weak Data in Regression, Wiley, New York, 1991.
[13] A. Dahlquist, A. Björk, N. Anderson, Numerical Methods, Prentice-Hall, Englewood Cliffs, NJ, 1974.
[14] W. Wagner, New vapor pressure measurements for argon and nitrogen and a new method for establishing rational vapor pressure equations, Cryogenics 13 (1973) 470–482.
[15] J. McGarry, Correlation and prediction of the vapor pressures of pure liquids over large pressure ranges, Ind. Eng. Chem. Process. Des. Dev. 22 (1983) 313–322.
[16] R.C. Weast (Ed.), Handbook of Chemistry and Physics, 56th ed., CRC Press, Ohio, 1975.
[17] J. Timmermans, Physico-Chemical Constants of Pure Organic Compounds, 2nd ed., Elsevier, New York, 1965.