Considering Error Propagation in Stepwise Polynomial Regression

Neima Brauner*

School of Engineering, Tel-Aviv University Israel, Tel-Aviv, 69978 Israel

Mordechai Shacham

Department of Chemical Engineering, Ben-Gurion University of the Negev, Beer-Sheva, 84105 Israel

The selection of an optimal regression model comprising linear combinations of various integer powers of an independent variable (explanatory variables) is considered. The optimal model is defined as the most accurate (minimal variance) stable model, where all parameter estimates of the orthogonalized explanatory variables are significantly different from zero. The potential causes that limit the number of terms that can be included in a stable regression model are investigated using two indicators, which measure signal-to-noise ratios in the variables. The truncation-to-noise ratio indicator is used to measure the extent of collinearity between the explanatory variables and the correlation-to-noise ratio indicator to evaluate the significance of the correlation between an explanatory variable and the dependent variable. It is shown that the number of terms that can be included in a stable polynomial model (and its accuracy) depend on the range and precision of the data, the rate of the error propagation during computations, and the algorithm used to calculate the regression parameters. It is demonstrated that it can often be advantageous to include nonconsecutive powers of the independent variable in an optimal polynomial model. An orthogonalized-variable-based stepwise regression procedure is presented, which enables identifying the optimal model in polynomial regression.

1. Introduction

Polynomials are frequently used for modeling and regression of thermophysical data, either as a complete empirical model (heat capacity, for example) or as a complement to theory-based models to represent deviations that are yet unexplained by the available theory (see, for example, Wagner's[1] equation for vapor pressure).

Because of the empirical nature of polynomial regression, it is difficult to determine a priori what the terms are that should be included in an optimal regression model. If the regression model contains either an insufficient number of terms or too many terms, inaccuracy and instability of the model may result. For instance, in published collections of regression models for physical and thermodynamic data, there is a tendency to set an absolute limit to the order of the polynomial that is fitted to a particular property. The same order of the polynomial is usually fitted to various data sets, irrespective of the range and precision of the available data. For example, in the collection of correlations of thermophysical properties,[2] the maximal polynomial order for solid, liquid, or gas heat capacity was set to four, without considering the large differences in the range and precision of the data available for the various phases.

Brauner and Shacham[3−5] have demonstrated some of the ill effects of including too many terms in a polynomial and other types of regression models. These effects include errors in the estimated values of the dependent variable, more severe errors in its derivatives with respect to the independent variables, and drastical change of the parameter estimates as a result of a small perturbation of the data (by removing or adding some data points, for example).

For high-precision data, higher order polynomials must be used for obtaining a regression model that predicts values within experimental error. In the chemical engineering community, there is a reluctance to use higher order polynomials because of the misconception that the highest order of the polynomial that can be used is limited by collinearity (for example, Lapidus[6] has set this limit to six unless orthogonal polynomials are used). Shacham and Brauner[7] and others have shown that this limit is imposed by near singularities when algorithms that are extremely sensitive to numerical error propagation are being used for calculating the regression model parameters.

There are several regression techniques, which possess low sensitivity to numerical error propagation, such as QR decomposition[8] and principal components regression, PCR.[9] If one of those techniques is used with precise data, instability caused by collinearity will not appear, even for very high order polynomials. However, when those numerically stable methods are applied to real life, noisy data it becomes apparent that numerical singularity is not the only cause of imprecision and instability in polynomial regression.[7] The order of the most accurate, stable polynomial is often limited by the noise level and range of the independent (and dependent) variables data.

In this paper, the requirements for a most accurate and stable, optimal polynomial regression model are investigated and specified. A regression algorithm based on orthogonalized variables (ROV) is presented. Indicators based on the signal-to-noise ratio in the orthogonalized variables are used to determine the highest order term that can be included in the regression model. A new, unified approach, which is based on perturbation

* To whom correspondence should be addressed. Phone: 972-3-6408127. Fax: 972-3-6429540. E-mail: brauner@eng.tau.ac.il.

10.1021/ie9901978 CCC: $18.00
Published on Web 10/07/1999

analysis for estimating the effect of propagated errors, regardless of their sources, is presented.

It can often be beneficial to use a polynomial regression model, which does not contain all the consecutive powers of the independent variable. To select the powers of the independent variable that should, or should not, be included in the regression model, a stepwise regression procedure can be employed. The ROV algorithm and the various indicators are combined to yield a stepwise regression procedure (SROV), which determines the optimal polynomial regression model for a particular set of data.

Two examples, which demonstrate the use of the various algorithms and indicators and emphasize their advantages over traditional statistical methods, are presented. The numerical calculations were carried out by the POLYMATH 4.0 (POLYMATH is copyrighted by M. Shacham and M. B. Cutlip, http://www.polymath-.software.com) and MATLAB 5.2 (MATLAB is a trademark of The Math Works, Inc., http://www.mathworks.com) packages. All the calculations were carried out on a PC using double-precision computation.

2. Basic Concepts

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 generalized polynomial regression model will be considered:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_n x_i^n + \epsilon_i \qquad (1)$$

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

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y} \qquad (2)$$

The rows of $\mathbf{X}$ are $\boldsymbol{x_i} = 1, x_i, ..., x_i^n$ and $\mathbf{X}^T\mathbf{X} = \mathbf{A}$ is the normal matrix. Another method to obtain the parameter values in eq 1 is by solving the following overdetermined system of equations:

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} \qquad (3)$$

using QR decomposition.[8] The ROV procedure, which will be presented in section 5, can also provide the parameter estimates. The various methods possess various levels of sensitivity to numerical error propagation, as will be demonstrated in a subsequent section.

A numerical indicator used most frequently to test for the quality of the fit is the sample variance, s^2, obtained by

$$s^2 = \frac{1}{\nu}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad (4)$$

Thus, the sample variance is the sum of squares of residuals divided by ν degrees of freedom ($\nu = N - (n + 1)$, where the number of parameters, $n + 1$, is subtracted from the number of data points, N). A smaller variance indicates a better fit of the model to the data. The standard error of the estimate, s, is a measure of variability around the line of regression. It is used to calculate the confidence interval ($\Delta\beta_j$) on a parameter value and is defined by

$$\beta_j = \hat{\beta}_j \pm \Delta\beta_j; \quad \Delta\beta_j \equiv (\nu, \alpha)s\sqrt{a_{jj}} \qquad (5)$$

where a_{jj} is the diagonal element in the $\boldsymbol{A}^{-1}$ matrix, $t(\nu, \alpha)$ is the statistical t distribution corresponding a desired confidence level, α. Clearly, if $\hat{\beta}_j$ is smaller (in its absolute value) than $\Delta\beta_j$(or $\Delta\beta_j/|\beta_j| > 1$), then the zero value is included inside the confidence interval, implying that there is no statistical justification to include the associated term in the regression model. Note that when the explanatory variables are strongly correlated, the individual confidence intervals will usually underestimate the uncertainty in the parameter estimates, as indicated by the confidence region. In this work, confidence intervals on parameter estimates of orthogonalized variables (no correlation between the variables) will be used as indicators.

An optimal model is defined as the most accurate (yields a minimal variance) stable model. Using statistical indicators, a stable model is defined as one where all the 95% confidence intervals are smaller than the respective parameter estimates for orthogonalized variables. In section 5, stability indicators, which are based on the signal-to-noise ratio, will be presented.

3. Collinearity and Error Propagation in Polynomial Regression

Collinearity between the explanatory variables often limits the number of terms that can be included in a stable and statistically valid polynomial model. Shacham and Brauner[7] investigated the various approaches to diagnose harmful levels of collinearity in polynomial regression and derived a criterion to measure the collinearity level. This criterion can be related to the mathematical definition of collinearity.

Following Gunst,[10] a collinearity is said to exist among the columns of $\mathbf{X} = [\mathbf{x}^0, \mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^n]$, 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

$$c_0\mathbf{x}^0 + c_1\mathbf{x}^1 + c_2\mathbf{x}^2 + ... + c_n\mathbf{x}^n = \boldsymbol{\Delta}; \quad \text{with } ||\boldsymbol{\Delta}|| < \eta \cdot ||\mathbf{c}|| \qquad (6)$$

where the notation $\mathbf{x}^j$ is used to indicate the vector whose elements are x_i^j.

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. Equation 6 can be divided by, say, $c_j (c_j \neq 0)$ to yield

$$c_{0,j}\mathbf{x}^0 + c_{1,j}\mathbf{x} + c_{2,j}\mathbf{x}^2 + ... + \mathbf{x}^j + ... c_{n,j}\mathbf{x}^n = \boldsymbol{\Delta}_j \qquad (7)$$

where $c_{k,j} = c_k/c_j$. $k = 1, 2, ..., n (c_{jj} \equiv 1)$. For a finite number of data points, the coefficients c can be obtained by regressing $\mathbf{x}^j$ on the remaining explanatory variables (or by orthogonal polynomials in the case of an infinite number of data points). The vector $\boldsymbol{\Delta}_j$ is the residual of this representation and is denoted as the "truncation error" (The term “truncation error” is used for $\boldsymbol{\Delta}_j$ because it represents the information lost (truncated) when $\mathbf{x}^j$ is removed from the model.). Since the values of the explanatory variables are subject to an error, the value of $\boldsymbol{\Delta}_j$ is also subject to an error, denoted by $\boldsymbol{\delta}_j$. Shacham and Brauner[7] suggested that in order for $\mathbf{x}^j$ to contain useful information, the norm of $\boldsymbol{\Delta}_j$ must be larger than the norm of the errors $\boldsymbol{\delta}_j$. Thus, the truncation-to-noise ratio (TNR) is defined by

$$\mathrm{TNR} = \frac{||\boldsymbol{\Delta}_j||}{||\boldsymbol{\delta}_j||} \tag{8}$$

When TNR $\gg$ 1, there is plenty of useful information in $\mathbf{x}^j$, thus, its inclusion in the model can be justified. A value of TNR close to 1, or smaller than 1, indicates that the information in $\mathbf{x}^j$ that is orthogonal to the other powers of $\mathbf{x}$ is mostly noise. Thus, including $\mathbf{x}^j$ in the model will render an unstable and statistically invalid model.

The error $\boldsymbol{\delta}_j$ can be calculated based on the definition of $\boldsymbol{\Delta}_j$ (eq 7). For error propagation calculation, two types of errors in the data must be separately considered. The first type is the experimental error caused by the limited precision of the measuring or control devices and due to reporting rounded values of the data. The second type is numerical error, which results from the limit on the number of digits carried by the computer. When the explanatory variables are functions of one or more independent variables (as in polynomial regression), the errors of the first type are carried over from the independent variables to their functions. In this case, the errors in the various terms are not independent, but correlated. On the other hand, numerical errors are uncorrelated. The rules of error propagation are different for correlated and uncorrelated errors and that makes the estimation of $\boldsymbol{\delta}_j$ rather cumbersome.

The propagated error can, however, be calculated using numerical perturbation. Using this technique, all the steps of the regression are carried out using two sets of the independent variable data: the original set, $\mathbf{x}$, and a perturbed set, $\mathbf{x}_e = \mathbf{x} + \boldsymbol{\eta}\cdot\boldsymbol{\delta}_j$, where the random $\boldsymbol{\eta}$ are scaled to yield the corresponding standard deviation of the error; thus $\eta_i \in [-5/3, 5/3]$ (for consideration of scaling $\boldsymbol{\eta}$, see, for example, Stewart.[11]) The norm of the difference between $\boldsymbol{\Delta}_j$ calculated using the original data and the perturbed data provides an estimate for $||\boldsymbol{\delta}_j||$.

In polynomial regression, collinearity and the consequent ill-conditioning of the problem can be substantially reduced by using certain transformations of the independent variable data. In this work, the *v*- and *z*-transformations[7] will be used. The *v*-transformation is defined by $v_i = x_i/x_{\max}$, where $x_{\max}$ is the largest x (in absolute value). This transformation yields a variable distribution in the region $v_{\min} \le v_i \le 1$. The *z*-transformation,

$$z_i = \frac{2x_i - x_{\max} - x_{\min}}{x_{\max} - x_{\min}} \tag{9}$$

yields a variable distribution in the range of $-1 < z_i \le 1$.

It is well-known that *v*-transformation (or no transformation) yields data that is much more sensitive to the harmful effects of collinearity than the *z*-transformation. The *v*-transformation and inversion of the normal matrix for regression (eq 2) are used to demonstrate the ability of the proposed techniques to diagnose collinearity.

4. Collinearity in Regression with Legendre Polynomials

It has long been recognized that the use of orthogonal polynomials can reduce considerably the ill effects of collinearity.[12] In this section the theoretical bound for the number of terms that can be used in polynomial regression, before the harmful effects of collinearity will appear in actual numerical computation, is determined. To that aim, we consider the independent variable to be continuous over the entire interval (of the *z*-transformation) [−1,1].

Legendre polynomials, $P_j(z)$, form a complete orthogonal set on [−1,1]. Furthermore, it can be proved that for a continuous function, $f(z)$ (satisfying the Dirichlet conditions), an expansion (as far as the desired degree of polynomial approximation) in Legendre polynomials gives the best least squares approximation. The coefficients for representing $f(z) = z^j$ in terms of $P_k(z)$, $z^j = \sum_{k=0}^{j} \alpha_k P_k(z)$ are tabulated.[13] The first six Legendre polynomials, scaled to yield the respective truncation error, Δ_j are given by

$$\begin{aligned}
\Delta_0(z) &= 1 \\
\Delta_1(z) &= z \\
\Delta_2(z) &= \frac{1}{3}(-1 + 3z^2) \\
\Delta_3(z) &= \frac{1}{5}(-3z + 5z^3) \\
\Delta_4(z) &= \frac{1}{35}(3 - 30z^2 + 35z^4) \\
\Delta_5(z) &= \frac{1}{63}(15z - 70z^3 + 63z^5) \\
\Delta_6(z) &= \frac{1}{231}(-5 + 105z^2 - 315z^4 + 231z^6)
\end{aligned} \tag{10}$$

Note that eq 10 defines the coefficients $c_{k,j}$ in eq 7. These polynomials can be integrated over the [−1,1] interval to yield the norm of the truncation error

$$||\boldsymbol{\Delta}_j|| = \sqrt{\frac{1}{2}\int_{-1}^{1} \Delta_j^{\,2}(z)\,\mathrm{d}z} \tag{11}$$

The propagated error in $\boldsymbol{\Delta}_j$ due to experimental error in the independent variable, $\boldsymbol{\delta}\mathbf{z}_c$ is

$$(\boldsymbol{\delta}_j)_c = \left|\frac{\mathrm{d}\boldsymbol{\Delta}_j}{\mathrm{d}z}\right|_c \boldsymbol{\delta}\mathbf{z}_c \tag{12}$$

Note that in calculating $(\boldsymbol{\delta}_j)_c$, the sign of the various terms composing $\boldsymbol{\Delta}_j$ is preserved, since the errors of the terms are correlated. On the other hand, when applying the error propagation formula for the numerical (uncorrelated) error $\boldsymbol{\delta}\mathbf{z}_u$), the sign of the various term is ignored; thus,

$$\begin{aligned}
(\boldsymbol{\delta}_j)_u = |c_{1,j}||\boldsymbol{\delta}\mathbf{z}_u| + |c_{2,j}||2\mathbf{z}\boldsymbol{\delta}\mathbf{z}_u| + \ldots + |j\mathbf{z}^{j-1}\cdot\boldsymbol{\delta}\mathbf{z}_u| + \\
\ldots + |c_{n,j}||n\mathbf{z}^{n-1}\cdot\boldsymbol{\delta}\mathbf{z}_u| = \left|\frac{\mathrm{d}\Delta_j}{\mathrm{d}z}\right|_u |\boldsymbol{\delta}\mathbf{z}_u
\end{aligned} \tag{13}$$

The estimation for $||\boldsymbol{\delta}_j||$ is obtained by combining the effects of $(\boldsymbol{\delta}_{jc})$ and $(\boldsymbol{\delta}_{ju})$

$$||\boldsymbol{\delta}_j|| = \left\{\frac{1}{2}\int_{-1}^{1}\left[\left|\frac{\mathrm{d}\boldsymbol{\Delta}_j}{\mathrm{d}z}\right|_c \left|\boldsymbol{\delta}z_c\right| + \left|\frac{\mathrm{d}\boldsymbol{\Delta}_j}{\mathrm{d}z}\right|_u \left|\boldsymbol{\delta}\mathbf{z}_u\right|\right]^2 \mathrm{d}z\right\}^{1/2} \tag{14}$$

Note that eqs 11−14 can be used for calculating $||\boldsymbol{\Delta}_j||$ and $||\boldsymbol{\delta}_j||$ for transformations other than the *z*-transformation, but the appropriate forms of the Legendre

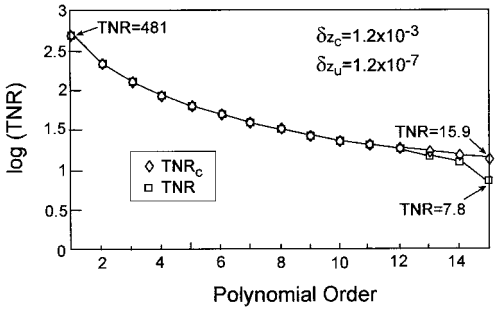

Figure 1. TNR versus polynomial order calculated using Legendre polynomials.

polynomials must be introduced into the equations and the limits on the integration must be changed.[7]

The effects of error propagation on the TNR are demonstrated in Figure 1. The figure shows TNRs versus the polynomial orders. The range of the independent variables was set to $x_{max}= 1000$ *and* $x_{min}= 500$. It was assumed that the data precision is three decimal digits, yielding an average correlated error of $\boldsymbol{\delta x_c} \sim 0.3$, and that single-precision computation is used with an average uncorrelated error of $\boldsymbol{\delta x}_u = 3 \times 10^{-5}$. The figure was prepared using *z*-transformation, but the results are invariant with the transformation. This is expected, since collinearity is a group phenomenon which is associated with a set of vectors and is invariant with linear transformation.

Two curves are shown. In the upper curve, only the propagation of the correlated (experimental) error is accounted for; in the second curve both the correlated and uncorrelated errors are considered. The two curves are practically identical up to the eleventh order polynomial, indicating that uncorrelated (numerical) error propagation has a negligible effect on the TNR. From this point on, the rate of decrease of TNR accelerates because of propagation of the numerical error. But, even so, the lowest value of TNR reached for the fifteenth order polynomial is 12.4, above the limit of TNR = 1, where the harmful effects of collinearity will prevail.

It is a common belief that the use of orthogonal polynomials completely eliminates the collinearity problem. Theoretically, orthogonal polynomials cannot be collinear, but with finite precision numerical computation, a value of a particular term may become small enough so that the numerical error dominates over its true value. This is the explanation for the continuous decrease in the value of the collinearity indicator, (TNR) with increasing order of the Legendre polynomial, in Figure 1. The trend of the curve in Figure 1 also indicates that the theoretical limit on the polynomial order (based on collinearity considerations alone) is very high. This leads to the popular misconception stated above.

5. Using Orthogonalized Explanatory Variables in Regression

For a finite number of data points, arbitrarily distributed on [−1,1], the use of orthogonal polynomials (e.g., Legendre, Chebyshev) may yield normal matrices with non-zero off-diagonal elements. Therefore, in practice, numerical orthogonalization of the explanatory variables, which yields a diagonal normal matrix irrespective of the variable spacing is required. It is to be noted that accounting for the errors propagated in the orthogonalization process is essential for detecting the harmful effects of collinearity.

Numerical orthogonalization of the explanatory variables is carried out by regressing one of the explanatory variables on the remaining ones (see, for example, pp 99−103, in Mandel[14]). The following implementation of this orthogonalization method is especially suitable for polynomial regression, where at every stage, an additional explanatory variable is added to the model (as in stepwise regression).

Let us define $\mathbf{x}^j{}_0 = \mathbf{x}^j$ (j = 1, 2, ..., n) N vectors of explanatory variables and $\mathbf{y}_0 = \mathbf{y}$ is the vector of the dependent variable. The subscript indicates that the values are for stage 0 of the ROV (regression using orthogonalized variables) procedure. For a model that contains a free parameter, the explanatory and dependent variables are first mean-centered; otherwise, no mean-centering is required.

At every stage of the regression process, the subsequent power of the independent variable is added to the model and the remaining (explanatory and dependent) variables are updated. At stage k, where all terms up to x^k have already been included in the model, the updated values are obtained by

$$\mathbf{y}_{k+1} = \mathbf{y}_k - \beta_k \mathbf{x}_k^k$$

where

$$\beta_k = \frac{(\mathbf{y}_k)^{\mathrm{T}} \mathbf{x}_k^k}{(\mathbf{x}_k^k)^{\mathrm{T}} \mathbf{x}_k^k} \tag{15}$$

is the coefficient of the kth orthogonal polynomial term and

$$\mathbf{x}_{k+1}^j = \mathbf{x}_k^j - \mathbf{x}_k^k \left[\frac{(\mathbf{x}_k^j)^{\mathrm{T}} \mathbf{x}_k^k}{(\mathbf{x}_k^k)^{\mathrm{T}} \mathbf{x}_k^k}\right]; \quad j = k+1, \quad k+2, ..., n \tag{16}$$

Using these equations, the regression progresses so that the variables already included in the model ($\mathbf{x}_1^1$, $\mathbf{x}_2^2$, ..., $\mathbf{x}_k^k$) are orthogonal to each other and $\mathbf{y}_k$ represents the residual of $\mathbf{y}$ which is orthogonal to this subset of explanatory variables.

To enable estimation of the errors propagated to $\mathbf{y}_k$ and $\mathbf{x}_{k+1}^j$, the regression is carried out, in parallel, using two data sets of data. The original set ($\mathbf{x}$, $\mathbf{y}$) and a perturbed set ($\mathbf{x}_e = \mathbf{x} + \boldsymbol{\eta}\cdot\delta\mathbf{x}$ and $\mathbf{y}_e = \mathbf{y} + \boldsymbol{\eta}\cdot\epsilon$) are used. Consequently, TNR for the subsequent variable to be included in the model $\mathbf{x}^{k+1}$ can be calculated from the following equation:

$$\mathrm{TNR}_{k+1} = \left[\frac{\mathbf{x}^{\mathrm{T}}\mathbf{x}}{(\mathbf{x}-\mathbf{x}_e)^{\mathrm{T}}(\mathbf{x}-\mathbf{x}_e)}\right]^{1/2}; \begin{bmatrix} \mathbf{x} \equiv x_{k+1}^{k+1} \\ \mathbf{x}_e \equiv (\mathbf{x}_e)_{k+1}^{k+1} \end{bmatrix} \tag{17}$$

For the regression model to be accurate and stable, both the numerator and the denominator in the expression for β_k (eq 15) must be accurate. If $\mathrm{TNR}_{k+1} \leq 1$, then the denominator in eq 15 contains mostly noise. To check the accuracy of the numerator, another indicator, the CNR, which measures the signal-to-noise ratio of the correlation $(\mathbf{y}_{k+1})^{\mathrm{T}}\mathbf{x}_{k+1}^{k+1}$ is employed. In polynomial regression, when the propagated error is calculated by numerical perturbation, CNR_{k+1} can be expressed as

$$\mathrm{CNR}_{k+1} = \frac{|(\mathbf{y})^{\mathrm{T}}\mathbf{x}|}{|\mathbf{x}^{\mathrm{T}}(\mathbf{y} - \mathbf{y}_{\mathrm{e}}| + |\mathbf{y}^{\mathrm{T}}(\mathbf{x} - \mathbf{x}_{\mathrm{e}})|} \qquad (18)$$

where

$$\mathbf{x} \equiv \mathbf{x}_{k+1}^{k+1} \qquad \mathbf{y} \equiv \mathbf{y}_{k+1}$$
$$\mathbf{x}_{\mathrm{e}} \equiv (\mathbf{x}_{\mathrm{e}})_{k+1}^{k+1} \qquad \mathbf{y}_{\mathrm{e}} \equiv (\mathbf{y}_{\mathrm{e}})_{k+1}$$

A value of $\mathrm{CNR}_{k+1} \gg 1$ implies that $\mathbf{x}_{k+1}$ can be safely added to the model. If $\mathrm{CNR}_{k+1} \leq 1$ then the numerator of the expression for β_{k+1} contains mostly noise and $\mathbf{x}_{k+1}$ should not be added to the model. The addition of higher polynomial terms to the model stops when ill effects of collinearity prevail, as indicated by $\mathrm{TNR}_{k+1} \leq 1$, or when $\mathrm{CNR}_{k+1} \leq 1$.

For a small number of data points the stability limit on the highest order polynomial may be set by degrees of freedom rather than data precision consideration. To detect such a situation, the variance is monitored during the ROV procedure. An increase of the variance with increasing the polynomial order may signal instability caused by severe reduction of the degrees of freedom. In such a case, the addition of higher order terms stops.

Example 1. Fitting Polynomials to Vapor Pressure Data of Nitrogen. Vapor pressure data are appropriate for studies related to polynomial regression because high-precision data are available, and for a sufficiently wide temperature range, high- (over tenth-) order polynomials may be needed for appropriate representation of the data. (There are, of course, more appropriate theory-based models to represent vapor pressure data. See, for example, McGarry.[15])

Selected information related to the nitrogen's vapor pressure data provided by Wagner[1] is shown in Table 1. There are 68 data points, spread in the interval between the triple point (63.148 K) and the critical point (126.2 K). With the *v*-transformation, this distribution yields $v_{\min} \cong 0.5$. Following the discussion related to the experimental errors in Wagner,[1] it is assumed that the data are accurate up to the decimal digits reported, yielding an average value of $\delta T = 3 \times 10^{-4}$ and $\epsilon = 3 \times 10^{-5}$.

Figure 2 shows the TNR values for polynomials up to the fifteenth-order for the vapor pressure data. The theoretical values of the TNR were calculated using Legendre polynomials (TNR_t), and compared with the values obtained in regression with numerically orthogonalized polynomials using the *z*-transform (TNR_z) and *v*-transform (TNR_v). It can be seen that the theoretical values obtained using the Legendre polynomials are almost identical to those obtained using ROV with the *z*-transform. The TNR values calculated using the *v*-transformation follow the same curve up to the eleventh-order polynomial. From this point on, the rate of reduction of TNR for the *v*-transformation accelerates and gets below TNR = 1 for the fifteenth-order polynomial. Thus, for the vapor pressure data, if ROV is used with the *z*-transformation, no harmful effects of collinearity are to be expected, even for a fifteenth-order polynomial. However, when using ROV with the *v*-transformation, collinearity caused by accelerated propagation of numerical error put the stability limit on the fourteenth-order polynomial.

To investigate the role of the regression algorithm on the appearance of the harmful effects of collinearity, the TNR values (up to the fifteenth-order polynomial) were calculated using three additional algorithms: QR decomposition, normal equation (eq 2) solved as a system of linear equations, and normal equation solved by first inverting the $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ matrix. These calculations were carried out using both *z*- and *v*-transformations.

The results of the calculations have shown that when the *z* -transformation is used, the regression algorithm used has no noticeable effects up to the fourteenth-order polynomial. The TNR values obtained by all algorithms are identical to those shown in Figure 2 (obtained using the ROV procedure). For the fifteenth-order polynomial, only the method which uses the inverse of the $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ matrix yields a TNR value which is significantly (3 times) smaller than the TNR obtained using the other methods.

With the *v*-transformation, Figure 3 shows that the effect of the regression algorithm used is much more pronounced. When the algorithm that solves the normal equation by inverting the $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ matrix is used, TNR is considerably reduced already for the sixth-order polynomial and it gets below TNR = 1 for the seventh-order polynomial. When the normal equations are solved as a set of linear equations, the same trend is indicated, but TNR gets below the threshold value of 1 for the seventh-order polynomial. The QR decomposition and the ROV yield identical TNR values up to the eleventh-

Table 1. Selected Information for Vapor Pressure Data (from Wagner, 1973)

	T (K)	*P* (bar)	*v*	*z*
minimal value	63.148	0.1252	0.50038	−1
maximal value	126.2	34.002	1	1
no. of data points	68			
avg. δT (K)	0.0003			
avg. δP (bar)	0.00003			
avg. δv	2.3772×10^{-6}			
avg. δz	9.5160×10^{-6}			

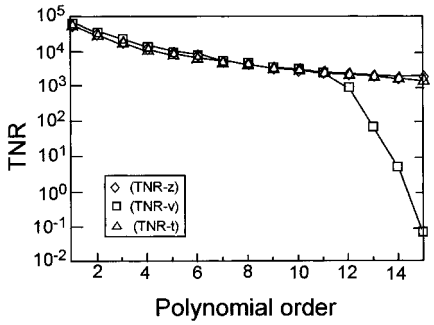

Figure 2. TNR versus polynomial order for nitrogen's vapor pressure data—comparison of the values obtained with Legendre polynomials (TNR-*t*), *v*-transformation (TNR-*v*), and *z*-transformation (TNR-*z*).

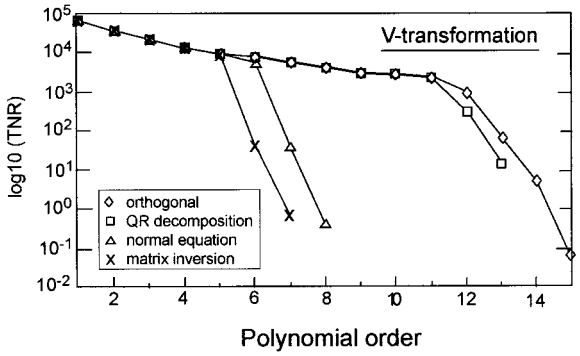

Figure 3. TNR values calculated for example 1 using four different regression algorithms with *v*-transformation.

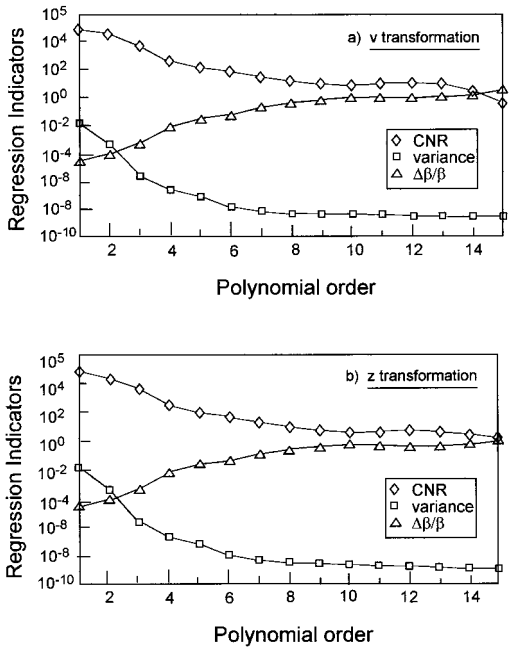

Figure 4. CNR, s^2, and $\Delta\beta_j/|\beta_j|$ for example 1 when (a) *v*-transformation and (b) *z*-transformation of the temperature are used.

order polynomial; from this point the decline rate of TNR values starts to be higher for the QR decomposition. The QR decomposition can be used only up to the thirteenth-order polynomial; after that the MATLAB program (used for this purpose) indicates a singular matrix and reduces the rank of the problem, essentially preventing the addition of higher order polynomial terms.

The above analysis suggests that comparison of the TNR values obtained with a particular data set using Legendre polynomials (or numerically orthogonalized polynomials with *z*-transformation) with those obtained using numerical orthogonalization with *v*-transformation can be a basis for determining the sensitivity of a particular regression algorithm to numerical error propagation. The TNR values obtained using the various transformations can be plotted versus the order of the polynomial. Comparing such a plot with Figure 3 can provide an estimate to the sensitivity of the tested algorithm to numerical error propagation (relative to ROV or the solution of the normal equation).

It should be emphasized that the theoretical analysis, using Legendre polynomials, is based on the assumption of an infinite number of data points. The numerical results of this example are based on a particular set of data with 69 data points and for polynomials up to the fifteenth-order. The use of a too small sample (with a high-order polynomial model) may lead to different results of no general value.

Collinearity considerations alone cannot determine the highest order of a polynomial that can be fitted to a particular set of data, since the signal-to-noise ratio of $\mathbf{y}^T\mathbf{x}$ (CNR, defined in eq 18) must also be considered. To demonstrate this point, the vapor pressure data of nitrogen were regressed with polynomial models up to the fifteenth degree. Normalized vapor pressure ($\pi_i = p_i/34.002$) was used as the dependent variable and *v*- or *z*-transformations of the temperature were used as the independent variable. The ROV procedure was used for regression.

Figure 4 shows CNR, s^2, and $\Delta\beta_j/|\beta_j|$ versus the

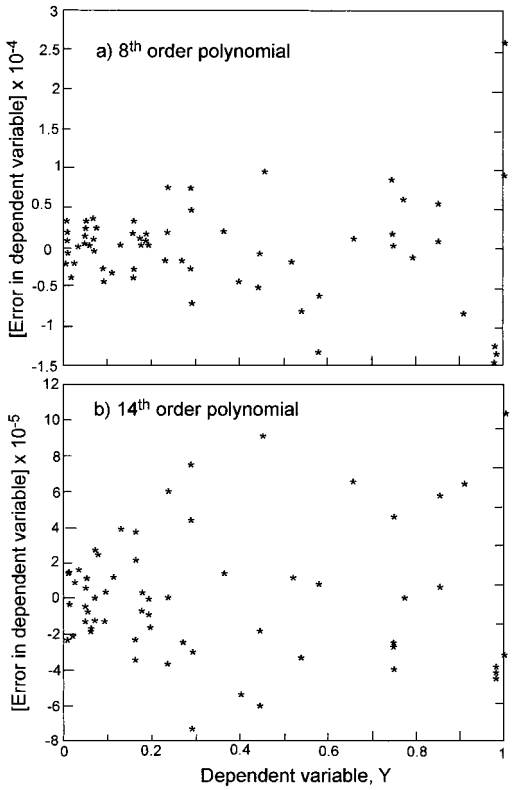

Figure 5. Residual plots for polynomial representation of the vapor pressure data, example 1.

polynomial order when the *v*-transformation is used. It can be seen that, at the beginning, the variance decreases sharply from $s^2 = 0.0136$ (for a linear model) to $s^2 = 3.63 \times 10^{-9}$ obtained with the eighth-order polynomial. From this point on, the decrease is much more moderate and s^2 reaches the value of 1.47×10^{-9} for the fifteenth-order polynomial.

The CNR value decreases from the initial value of $CNR_1 = 7.1 \times 10^4$ to $CNR_{14} = 1.37$. For the fifteenth-order polynomial $CNR_{15} = 0.199$, indicating that for the fifteenth term of the model, the noise in $\mathbf{y}^T\mathbf{x}$ is dominant; thus, this term should not be included in the model. (We may recall that TNR_{15} was also smaller than 1; see Figure 3). The value $\Delta\beta_j/|\beta_j|$ is a mirror image of the CNR values. For the first-order polynomial, $\Delta\beta_1/|\beta_1| = 3.14 \times 10^{-5}$ and it increases to 0.6799 for the fourteenth-order polynomial. For the fifteenth-order polynomial $\Delta\beta_{15}/|\beta_{15}| = 1.267$, indicating that the confidence interval on the respective parameter is larger than the parameter value itself. In this case, all the indicators (TNR, CNR, and $\Delta\beta_j/|\beta_j|$) consistently signal that the fourteenth-order polynomial is the highest order stable and statistically valid polynomial.

It is interesting to note the similarity of the CNR and the $\Delta\beta_j/|\beta_j|$ curves in Figure 4. The calculation of confidence intervals is based on statistical principles, while the calculation of CNR is based on error propagation considerations; yet, they both show the very same trend.

The moderate decrease of the variance starting with the eighth-order polynomial implies that the improvement of the data representation using higher order polynomials is not very significant. Figure 5 shows the residual plots for the eighth- and fourteenth-order polynomials. It can be seen that for the eighth order polynomial, the error is the largest near the critical

Table 2. Selected Information for Heat Capacity Data (from Timmermans, 1965)

	T (K)	C_p (cal/g·K)	v	z	C_p (norm.)
minimal value	14.17	0.0271	0.166981	−1	0.079941
maximal value	84.86	0.339	1	1	1
no. of data points	19				
avg. δT	0.03				
avg. δC_p	0.0003				
avg. δv	0.000354				
avg. δz	0.000849				
avg δC_p (norm.)	0.000885				

point, where $\pi = p/p_c \sim 1$. Increasing the polynomial order to 14 essentially reduces the errors in this region to the error level achieved in the rest of the interval.

6. Selection of the Optimal Regression Model

The polynomial model, which consists of sequential powers of the independent variable, is not necessarily the optimal model. The ROV procedure can be easily modified, so that at every stage of the process the explanatory variable (the particular power of the independent variable), which causes the maximal reduction of the variance, is included in the model.

The strength of the linear correlation between an explanatory variable $\mathbf{x}_k^j$ and the dependent variable $\mathbf{y}_k$ is measured by

$$YX_j = (\mathbf{y}_k)^T \mathbf{x}_k^j \tag{19}$$

where $\mathbf{y_k}$ and $\mathbf{x}_k^j$ are normalized to a unit length. The value of $|YX_j|$ is in the range [0, 1]. In a case of a perfect correlation between $\mathbf{y}_k$ and $\mathbf{x}_k^j$ ($\mathbf{y}_k$ is aligned in the $\mathbf{x}_k^j$ direction), $|YX_j| = 1$. In case $\mathbf{y}_k$ is unaffected by $\mathbf{x}_k^j$ (the two vectors are orthogonal), $YX_j = 0$. The inclusion of $\mathbf{x}_k^p$ associated with max $\{YX_j\}$ will affect the maximal reduction of the variance of the regression model. Further discussion and graphical interpretation of the YX_j values will be provided in connection with example 2. This indicator will be used to select the next variable to be included in a regression model at each stage of the stepwise regression (SROV).

If $TNR_j \leq 1$ or $CNR_j \leq 1$ for all the variables not included in the model, the addition of new variables to the model stops. The model selected this way is often the optimal model. To verify that the model obtained is indeed the optimal one, an additional phase of rotation of the order in which the explanatory (polynomial) terms are added to the model is carried out, so that each polynomial term in turn is selected as the last one to be included in the model.

The selection of an optimal regression model is explained in more detail and demonstrated in several examples by Shacham and Brauner.[16] The following example also demonstrates the advantages of the optimal polynomial model over a consecutive-power polynomial model.

Example 2. Correlation of Heat Capacity Data of Solid Propylene. Heat capacity versus temperature data are usually correlated by polynomials. Daubert and Danner,[2] for example, used a fourth-order polynomial to correlate heat capacity (C_P) data versus temperature for solid propylene, as published by Timmermans.[17] Selected information from Timmermans' data is shown in Table 2. Only the first 19 data points, out of 20 provided by Timmermans,[17] were used, since the last data point (nearly at the melting point) turned out to be very inaccurate because of premelting.

Polynomials of various orders were fitted to the data using v- and z--transformations for the independent variable (temperature) data. The dependent variable (C_P) was normalized by dividing it by the maximal value of C_P.

In Figure 6, mean-centered and scaled (to unit length) values of various powers of z are plotted versus mean-centered and scaled values of $C_P(\mathbf{y})$ at the various stages of the SROV procedure. The respective values of YX_j and CNR_j are also shown.

At the zeroth stage, it can be seen that the data points for $\mathbf{z}^1$ are aligned almost perfectly along the straight line with a slope of 1, which represents a perfect correlation. The respective YX_1 value is 0.995. The shape of the curve of $\mathbf{z}^3$ is similar to that of $\mathbf{z}^1$, but there is considerably larger curvature. Consequently, the value of YX_3 (0.945) is smaller than the value of YX_1. At this stage, the linear correlation between $\mathbf{y}$ and $\mathbf{z}^2$ or $\mathbf{z}^3$ is weak. This is indicated by very small absolute values of YX_2 (=0.0997) and YX_4 (=0.0903). However, at this stage, all TNR_j and CNR_j are much greater than 1; thus, stability considerations do not exclude inclusion of any one of the variables in the regression model.

Because of its highest YX_j, $\mathbf{z}^1$ is entered into the model, and in Figure 6, the updated values of $\mathbf{z}^2$ and $\mathbf{z}^3$ are plotted versus the updated values of $\mathbf{y}$, as obtained at stage 1. At this stage, there is already a considerable spread of the data points, but the points of $\mathbf{z}^2$ line up nicely along the straight line of a slope −1, with $YX_2 = -0.832$. The trend of the points representing $\mathbf{z}^3$ is not well-defined and the value of YX_3 is 0.497. Thus, $\mathbf{z}^2$, which was nearly orthogonal to $\mathbf{y}$ at the zeroth stage, becomes strongly correlated with the residual of $\mathbf{y}$, which is orthogonal to and is unexplained by $\mathbf{z}^1$. The linear correlation between $\mathbf{z}^3$ (which was nearly collinear to $\mathbf{z}^1$ at zeroth stage) and the residual of $\mathbf{y}$ is much weaker at this stage. It is interesting to note that CNR_3 was reduced by more than an order of magnitude after the information which is collinear to $\mathbf{z}^1$ has been subtracted from $\mathbf{z}^3$ and $\mathbf{y}$, while the value of CNR_2 has actually increased. Thus, at stage 1, $\mathbf{z}^2$ is added to the basis.

At stages 2 and 3, the situation described for stage 1 is repeated. The variable, which was nearly orthogonal to the variable that entered the basis in the previous stage and exhibited a weak correlation with $\mathbf{y}$, turned to be the one which is most strongly correlated with $\mathbf{y}$. Thus, $\mathbf{z}^3$ is added to the model at stage 2 and $\mathbf{z}^4$ at stage 3.

At stage 4, Figure 6 shows the updated values of $\mathbf{z}^5$ and $\mathbf{z}^8$ versus the updated values of $\boldsymbol{y}$. In spite of the increased spread of the data, the trend of the $\mathbf{z}^8$ data points, following the +1 slope line, is clearly distinguishable. In comparison, the $\mathbf{z}^5$ values are spread almost randomly with $|YX_5| = 0.162$. The CNR value is also much larger for $\mathbf{z}^8$ then for $\mathbf{z}^5$. Thus, at this stage, $\mathbf{z}^8$ is added to the model and not $\mathbf{z}^5$.

After including $\mathbf{z}^8$ in the model, stability considerations (CNR and confidence interval values) show that no more variables can be added.

Table 3 shows the regression results for third-, fourth-, and fifth-order polynomials and the optimal polynomial model for the v-transformation. The results include the parameter estimates, $\Delta\beta/|\beta|$, s^2, and sum of squares of errors. It can be seen that, for the third-order

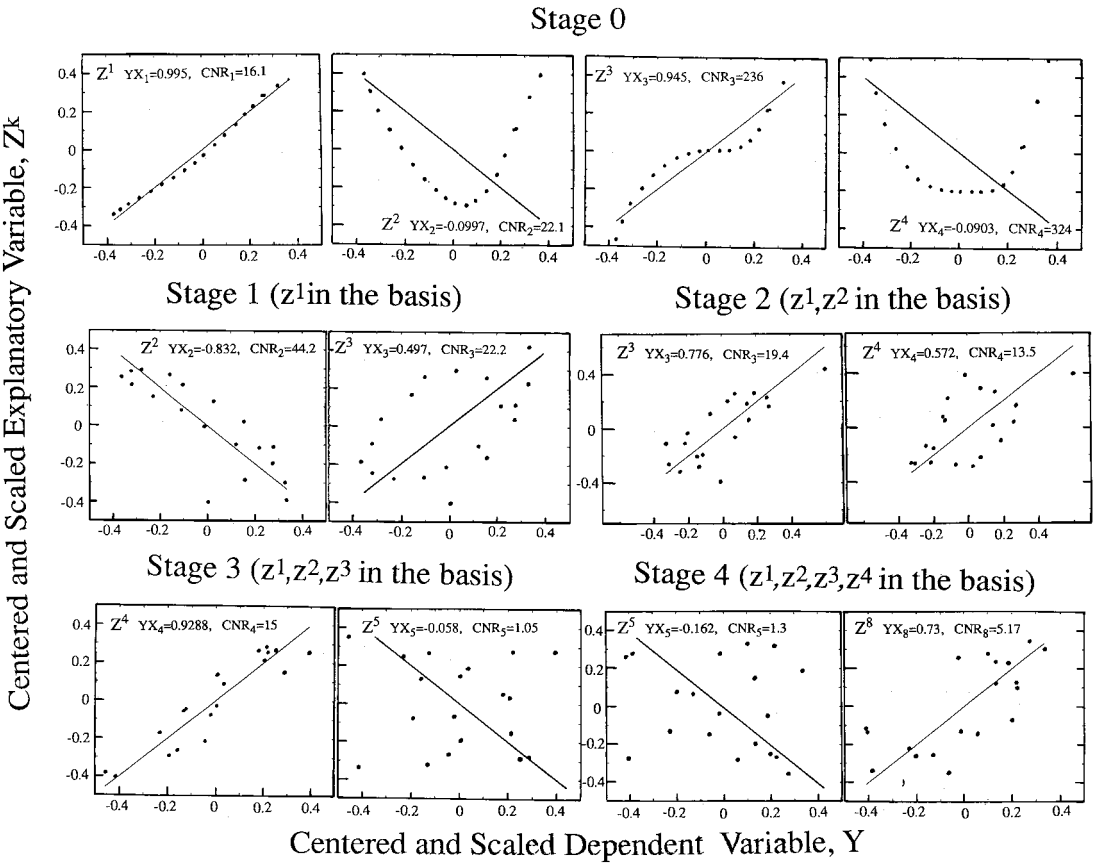

Figure 6. **y** vs z^k at various stages of the SROV procedure, example 2.

Table 3. Regression Results for Example 2 (*v*-Transformation Used)

	third-order polynomial		fourth-order polynomial		fifth-order polynomial		optimal	
	value	$\delta\beta/\lvert\beta\rvert$	value	$\delta\beta/\lvert\beta\rvert$	value	$\delta\beta/\lvert\beta\rvert$	value	$\delta\beta/\lvert\beta\rvert$
$\beta 0$	−0.287 18	−0.216 03	−0.075 35	−0.722 87	−0.043 53	−2.962 65	0.2177 18	0.749 626
$\beta 1$	2.404 99	0.168 4	0.371 259	1.332 224	−0.014 75	−101.537	−3.627 52	−0.595 4
$\beta 2$	−2.120 93	−0.363 16	4.250 54	0.352 661	5.929 17	1.065 023	24.254 2	0.436 118
$\beta 3$	0.984 937	0.447 846	−6.956 02	−0.264 38	−10.277 1	−1.192 99	−52.471 3	−0.453 11
$\beta 4$			3.407 88	0.230 554	6.444 12	1.723 09	46.800 5	0.481 918
$\beta 5$					−1.040 52	−3.647 32		
$\beta 6$							−26.510 7	−0.522 14
$\beta 7$							12.338 2	0.468 572
s^2	0.000129		1.92×10^{-5}		2.01×10^{-5}		9.07×10^{-6}	

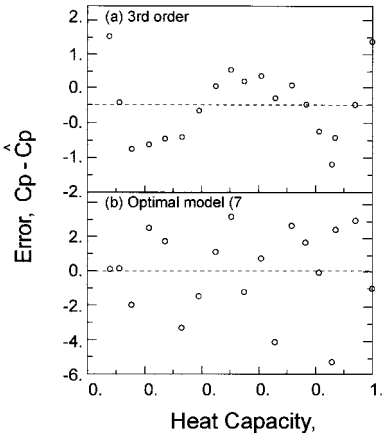

Figure 7. Residual plots for polynomial representation of the propylene heat capacity (*v*-transformation), example 2.

polynomial, all the $\Delta\beta/\lvert\beta\rvert$ values are smaller than 1; thus, significantly different from zero. However, the residual plot for this polynomial (shown in Figure 7a) shows a clear trend indicating that the third order polynomial is unsatisfactory for representing this data. For the fourth-order polynomial (as recommended by Daubert and Danner[2]), the coefficient β_1 is insignificant since $\Delta\beta/\lvert\beta\rvert = 1.33$, and for the fifth-order polynomial, all the coefficients are insignificant. The optimal model, which was found using the SROV procedure, contains all the powers of v, up to v^7 except v^5. In this model, all the $\Delta\beta_j/\lvert\beta_j\rvert$ values are smaller than one and the residual plot (Figure 7b) shows a random distribution. Thus, this seven-parameter model is statistically valid. Its s^2 is 14 times smaller than the s^2 of the third-order polynomial and two times smaller than the s^2 of the fourth- or fifth-order polynomial.

Table 4 shows the regression results for the fourth-order polynomial and the optimal polynomial model obtained with the *z*-transformation. In this case, all the coefficients of the fourth-order model are significant, but the use of the optimal model (consists of z, z^2, z^3, z^4 and z^8) enables further reduction of s^2 by a factor of 2. The s^2 value for the optimal model with *z*-transformation is also slightly lower than the s^2 value for the optimal model with *v*-transformation.

Table 4. Regression Results for Example 2 (*z*-Transformation Used)

	fourth-order polynomial		optimal	
	value	$\delta\beta/\lvert\beta\rvert$	value	$\delta\beta/\lvert\beta\rvert$
$\beta 0$	0.601 582	0.007 392	0.598 495	0.004 572
$\beta 1$	0.389 34	0.023 44	0.390 16	0.015 9977
$\beta 2$	−0.167 28	−0.141 08	−0.116 73	−0.072 7
$\beta 3$	0.072 1	0.171 845	0.070 882	0.119 422
$\beta 4$	0.102 561	0.230 594		
$\beta 8$			0.058 788	0.151 108
s^2	1.92×10^{-5}		8.95×10^{-6}	

Thus, restricting the polynomial model to consecutive powers of the independent variable may prevent obtaining a statistically valid model because the inclusion of sufficient consecutive terms for obtaining a random residual distribution may render an unstable model with some of the parameter values not significantly different from zero. Using a stepwise regression procedure (such as the SROV procedure) to select the particular terms that should be included in the model yields a stable and statistically valid model that also represents the data more accurately.

7. Conclusions

It has been shown that the range and precision of the data, error propagated during computations and the algorithm used to find the model parameters, have equally important roles in determining the maximum number of terms that can be included in the most accurate, stable polynomial model.

In many cases, the limit on the number of polynomial terms that can be included in a stable model is imposed by the harmful effects of collinearity. Collinearity effects, caused by error propagation, can be measured by the TNR indicator. For a particular polynomial, representing a particular set of data, the TNR attains maximal values (indicating minimal effects of collinearity) when there are an infinite number of data points and the regression is carried out with orthogonal (Legendre) polynomials. For a finite number of data points, using *z*-transformation and numerical orthogonalization, the obtained TNR values are very close to the theoretical bound, irrespective of the regression algorithm used. With the *v* -transformation and a regression algorithm which requires calculation of the normal matrix, the TNR values decline and may reach a value lower than 1 already for relatively low-order polynomials. In this case, collinearity puts a limit on the maximal polynomial order, which may be lower than that required for accurately representing the dependent variable data. The QR decompositions, or the ROV procedures, are insensitive to error propagation and yield TNR values closer to the theoretical bound obtained using Legendre polynomials. Thus, the use of these methods for obtaining the parameter values allows more accurate and stable representation of the data. If the sensitivity of the regression algorithm used for error propagation is not known, it is always recommended to verify results obtained using *v* -transformation (or no transformation) with those obtained using the *z* -transformation, which has much lower sensitivity to the harmful effects of collinearity.

Often, the maximal order of polynomials is limited by the combined effects of propagated errors in the explanatory and the dependent variables. The combined effects of these propagated errors is measured by the CNR indicator. This indicator, which is based on error propagation considerations, shows the very same trend as the indicator $\Delta\beta_j/|\beta_j|$. The latter is based on statistical considerations. An important advantage of the CNR indicator is that because it is based on mathematical and numerical principles (rather than statistical principles), it is not subject to the strict error distribution assumptions as the statistical indicator. Furthermore, examining the error terms included in the CNR can indicate whether the dominant errors are due to the independent variables or the dependent variables. The action needed to improve the accuracy of the model can be decided upon diagnosing the variable that introduces the dominant error.

It was shown that restricting the polynomial model to consecutive powers of the independent variable may yield a much less accurate model than the one that can be obtained without this restriction. The SROV procedure can help in identifying the optimal polynomial model.

Literature Cited

(1) Wagner, W. New Vapor Pressure Measurements for Argon and Nitrogen and a New Method for Establishing Rational Vapor Pressure Equations. *Cryogenics* **1973**, *13*, 470−482.

(2) Daubert, T. E.; Danner, R. P. *Physical and Thermodynamic Properties of Pure Chemicals: Data Compilation*; Hemisphere Publishing Co.: New York, 1989.

(3) Brauner, N.; Shacham, M. Considering Numerical Error Propagation in Modeling and Regression of Data. Proceedings of the 1998 ASEE Annual Conference, Seattle, WA, June 27−July 1, 1998.

(4) Brauner, N.; Shacham, M. Identifying and Removing Sources of Imprecision in Polynomial Regression. *J. Math. Comput. Simul.* **1998**, *48* (1), 77−93.

(5) Brauner, N.; Shacham, M. Role of Range and Precision of the Independent Variable in Regression of Data. *AIChE J.* **1998**, *44* (3), 603−611.

(6) Lapidus, L. *Digital Computation for Chemical Engineers*; McGraw Hill: New York, 1962.

(7) Shacham, M.; Brauner, N. Minimizing the Effects of Collinearity in Polynomial Regression. *Ind. Eng. Chem. Res.* **1997**, *36* (10), 4405−4412.

(8) Press,W. H., Teukolsky, S. A.; Vetterling, W. T., Flanerry, B. F. *Numerical Recipes in FORTRAN, The Art of Scientific Computing*, 2nd ed.; Cambridge University Press: Cambridge, 1992.

(9) Jackson, J. E. *A User Guide to Principal Components*; John Wiley: New York, 1991.

(10) Gunst, R. F. Toward a Balanced Assessment of Collinearity Diagnostics. *Am. Stat.* **1984**, *38* (2), 79−82.

(11) Stewart, G. W. Collinearity and Least Squares Regression. *Stat. Sci.* **1987**, *2* (1), 68−100.

(12) Seber, G. A. F. *Linear Regression Analysis*; Wiley: New York, 1977.

(13) Abramovitz, M.; Stegun, I. A. *Handbook of Mathematical Functions*, Dover Publications: New York, 1972.

(14) Mandel, J. *Evaluation and Control of Measurements, Quality and Reliability*; Marcel Dekker: New York, 1991.

(15) McGarry, J. Correlation and Prediction of the Vapor Pressures of Pure Liquids over Large Pressure Ranges. *Ind. Eng. Chem. Process. Des. Dev.* **1983**, *22*, 313−322.

(16) Shacham, M.; Brauner, N. Considering Precision of Experimental Data in Construction of Optimal Regression Models. *Chem. Eng. Process.* **1999**, *38* (4−6), 477−486.

(17) Timmermans, J. *Physico Chemical Constants of Pure Organic Compounds*; 2nd. ed., Elsevier, New York, 1965.

Received for review March 19, 1999
Revised manuscript received July 20, 1999

Resubmitted for review August 2, 1999

IE9901978