Selection of Simplified Models: II. Development of a Model Selection Criterion Based on Mean-Squared-Error

Shaohua Wu, K. B. McAuley and T. J. Harris*

Department of Chemical Engineering, Queen's University, 19 Division Street, Kingston, ON Canada K7L 3N6

Abstract

Simplified models (SMs) with a reduced set of parameters are used in many practical situations, especially when the available data for parameter estimation are limited. A variety of candidate models are often considered during the model formulation, simplification and parameter estimation processes. We propose a new criterion to help modellers select the best SM, so that predictions with lowest expected mean-squared-error can be obtained. The effectiveness of the proposed criterion for selecting simplified nonlinear univariate and multivariate models is demonstrated using Monte Carlo simulations and is compared with the effectiveness of the Bayesian Information Criterion (BIC).

Key Words: simplified models, mean-squared-prediction-error, phenomenological models, model selection criteria

1. Introduction

A mathematical model is a representation, in mathematical terms, of certain aspects of a nonmathematical system (Aris, 1999). In science and engineering, mathematical modelling plays an important role, and models are used for simulating, designing, controlling and optimizing industrial production processes. In many modelling situations in chemical

* Author to whom correspondence may be addressed. E-mail address: tom.harris@chee.queensu.ca.

engineering, modellers have sufficient scientific knowledge to derive complex phenomenological models, which can be expected to match the underlying behaviour of the process very well. Unfortunately, it is often too difficult or costly to obtain enough good data to reliably estimate all of the unknown model parameters (e.g., Perregaard, 1993; Bagajewicz and Cabrera, 2003; Yoshida et al., 2003; Lv et al., 2004; Maria, 2004, 2006; Mchaweh et al., 2004; Chang et al., 2005; Romdhane and Tizaoui, 2005; Wang et al., 2007). For complex models with many parameters, the resulting parameter estimates and model predictions may exhibit high variability, especially when the data available are limited (e.g., the number of data points is small, measurements are noisy, the range of input-variable settings is small, and/or experimental designs are highly correlated) (Wu et al., 2007). The decisions made using these models (or their parameter estimates) may be unreliable. As a result, it is important to avoid estimating too many model parameters using limited data.

Because of the difficulties associated with formulating complex models and with obtaining good estimates for all of the unknown parameters, engineers often use simplified models (SMs) that are known to be structurally imperfect. A variety of candidate SMs can be obtained by making different assumptions during the model formulation and simplification process. There are many reasons to choose a SM with fewer parameters and terms than the truly-structured or extended model (EM) (Zhang, 1997). The practical advantages of a parsimonious model often overshadow concerns over the correctness of the model structure. When the available data are not informative, SMs can be expected to give better predictions with lower mean-squared-error than the EM (Rao, 1971; Hocking 1976; Wu et al., 2007).

When experimental data are insufficient to support the use of complex models, modellers must make decisions about model simplification. They need to know which terms and parameters to include, which parameters to fix at nominal values, and which terms to leave

out, so that they can obtain the best possible predictions using the data that they possess along with their scientific and engineering knowledge. Many different strategies have been developed for selecting appropriate SMs.

Model-Selection Criteria (MSC) have been studied and used for model selection since the Akaike Information Criterion (AIC) was proposed in 1973 (e.g., Akaike (1973), Linhart and Zucchini (1986), McQuarrie and Tsai (1998), Rao and Wu (2001), Burnham and Anderson (2002), Konishi and Kitagawa (2008)). When using a MSC, such as the AIC, the Final Prediction Error Criterion (FPE), or Mallow's C_p, the criterion value for different candidate models is calculated directly from the model equations and the residuals that are obtained during parameter estimation. The candidate model with the lowest criterion value is selected as the best model. MSC are simple to use because no numerical optimization is required beyond the parameter estimation step. In the first article in this series (Wu et al., 2009), we compared nine commonly-used MSC for their performance and tendencies when selecting SMs when the number of data points is small and experimental designs are correlated. We showed that the expected mean-squared-error provides a convenient theoretical means for analyzing the relative tendencies of these different MSC.

In this article, we develop a new MSC aimed at selecting the SM with the lowest expected mean-squared-error (MSE) for model predictions made at the design points. This new MSC explicitly accounts for bias due to imperfect model structure and for variance in model parameters and predictions arising from noisy data.

It is well known that removing parameters from a truly-structured EM will introduce bias, but may decrease variance in model predictions (Rao, 1971; Hocking 1976). Use of the MSE for selecting appropriate SMs has been studied by Linhart and Zucchini (1986) and Wu et al. (2007). Linhart and Zucchini (1986) proposed a hypothesis-test approach to compare two

nested models and to select the one with lower mean-squared-prediction-error. Wu et al. (2007) summarized the many quantitative and qualitative results in the literature concerned with using and selecting SMs. A confidence-interval approach was then developed to assess the uncertainty associated with whether a SM or the EM will provide lower-MSE model predictions at the design points used for parameter estimation. It was shown that, when SMs are preferable due to limited data, decisions concerning whether the EM or SM will give better predictions are very uncertain.

One short-coming of the approaches proposed by Linhart and Zucchini (1986) and Wu et al. (2007) is that they can only be used for comparing two nested models, where the SM is a simplified version of the more complex EM. However, in many practical situations, modellers often consider a set of candidate SMs, which may or may not be nested with each other. In the current article, a model-selection criterion is proposed for selecting the best model (with the lowest expected MSE for predictions) from a group of candidate models that includes the EM and several SMs. This criterion is developed using univariate linear models, and is then extended for selection of univariate nonlinear models. The performance of the proposed model selection criterion is demonstrated using Monte Carlo simulations and is compared with the performance of the Bayesian Information Criterion (BIC). It is also shown that the proposed criterion is effective for selecting multivariate nonlinear models when the noise variance-covariance matrix is known. Difficulties associated with selecting simplified multivariate models when the noise variance-covariance matrix is unknown are discussed.

2. Development of MSE-Based Model-Selection Criterion

Consider a truly-structured EM that can be described by the following univariate linear model

$$\begin{aligned} Y &= X\beta + \varepsilon \\ &= X_1\beta_1 + X_2\beta_2 + \varepsilon \end{aligned} \tag{1}$$

where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$, $\varepsilon \in \mathbb{R}^n$, $X_1 \in \mathbb{R}^{n\times p_1}$, $\beta_1 \in \mathbb{R}^{p_1}$, $X_2 \in \mathbb{R}^{n\times(p-p_1)}$ and $\beta_2 \in \mathbb{R}^{p-p_1}$. In this model, the noise ε is additive, and the noise-free response of the process is $Y_{true} = X\beta$. Also, assume that (Beck and Arnold, 1977): 1) input settings X_1 and X_2 are deterministic with full column rank; and 2) the stochastic component ε is mean-zero and uncorrelated with constant variance σ^2. A particular simplified version of the EM is of the form

$$Y = X_1\beta_1 + e \tag{2}$$

where $e = X_2\beta_2 + \varepsilon$ incorporates the stochastic component combined with any model mismatch. The SM is nested within the EM.

In the current article, we are interested in selecting the best simplified model (SM) from a set of candidate models to obtain the lowest total mean-squared-error (MSE) for model predictions. The MSE is defined as the expected squared difference between the model prediction, $\hat{Y}$, and the noise-free response of the process, Y_{true} (Rice, 1995). For a column vector of predictions $\hat{Y}$ obtained using a candidate model, the total MSE is:

$$\begin{aligned} MSE(\hat{Y}) &= E\left(\left(\hat{Y} - Y_{true}\right)^T\left(\hat{Y} - Y_{true}\right)\right) \\ &= \left(E(\hat{Y}) - Y_{true}\right)^T\left(E(\hat{Y}) - Y_{true}\right) + tr\left(Cov(\hat{Y})\right) \end{aligned} \tag{3}$$

where $E(\cdot)$, $Cov(\cdot)$, and $tr(\cdot)$ denote the expected value, variance-covariance matrix and trace, respectively. The second line in Eqn. (3) shows that MSE is equal to the squared bias ($\left(E(\hat{Y}) - Y_{true}\right)^T\left(E(\hat{Y}) - Y_{true}\right)$) plus the total variance ($tr\left(Cov(\hat{Y})\right)$) of the model predictions (Rice, 1995). As a result, MSE, which accounts for both bias and variance, is an appropriate criterion for analyzing simplified or misspecified models.

When unknown parameters in the EM (Eqn. (1)) and SM (Eqn. (2)) are estimated using

ordinary least-squares (OLS), the expected total MSE for predictions is (Beck and Arnold, 1977)

$$MSE_E = \sigma^2 p$$
$$MSE_S = \sigma^2 p_1 + \beta_2^T X_2^T (I_n - P_1) X_2 \beta_2 \quad (4)$$

where the subscripts "*E*" and "*S*" indicate the use of the EM and the SM, respectively.

In our previous work (Wu et al., 2007), we developed a strategy to determine whether the SM or the EM is expected to give predictions with lower MSE (at the design points used for parameter estimation). This strategy relies on a critical ratio R_C, which is defined as

$$R_C = \frac{\beta_2^T X_2^T (I_n - P_1) X_2 \beta_2}{(p - p_1)\sigma^2} \quad (5)$$

where $P_1 = X_1 (X_1^T X_1)^{-1} X_1^T$. The numerator of R_C is the squared bias introduced by removing parameters associated with X_2 from the model, and the denominator is the variance reduction (due to fewer parameters being estimated) when a particular SM is used rather than the EM. As a result,

$$R_C < 1 \quad (6)$$

is a necessary and sufficient condition for $MSE_S < MSE_E$, which implies that the SM is preferable to the EM for making predictions. This critical ratio has also been used to compare the tendencies of various MSC that are commonly used for selecting SMs (Wu et al., 2009).

The value of R_C depends on unknown β_2 and σ^2. Fitting the EM using OLS provides $\hat{\beta}_2$ and s_E^2, which are unbiased estimators of β_2 and σ^2. Therefore, an estimator of R_C can be obtained as

$$\hat{R}_C = \frac{\hat{\beta}_2^T X_2^T (I_n - P_1) X_2 \hat{\beta}_2}{(p - p_1) s_E^2} = \frac{(SSE_S - SSE_E)/(p - p_1)}{SSE_E/(n - p)} \quad (7)$$

where "SSE" denotes the sum of squared residuals. Note that the calculation of $\hat{R}_C$ requires knowledge (or an assumption) about the form of the truly-structured EM. Given the additional

assumption that ε is normally distributed, $\hat{R}_C$ in Eqn. (7) is a likelihood ratio statistic (Wang and Chow, 1994), which follows a noncentral F distribution (Montgomery et al., 2001) with $(p - p_1)$ and $(n - p)$ degrees of freedom, and noncentrality parameter:

$$\lambda = (p - p_1)R_C \tag{8}$$

$\hat{R}_C$ given in Eqn. (7) is also the statistic of a partial F test for testing the hypothesis $H_0: \beta_2 = 0$ (Montgomery et al., 2001).

In situations where σ^2 is known from prior information about the variability of the response variable, an estimator for R_C can be obtained from:

$$\hat{R}_C = \frac{\hat{\beta}_2^T X_2^T (I_n - P_1) X_2 \hat{\beta}_2}{(p - p_1)\sigma^2} = \frac{(SSE_S - SSE_E)/(p - p_1)}{\sigma^2} \tag{9}$$

where $(p - p_1)\hat{R}_C$ follows a noncentral χ^2 distribution (Montgomery et al., 2001) with $(p - p_1)$ degrees of freedom and the noncentrality parameter λ from Eqn. (8).

From Eqns. (4) and (5), the reduction in total MSE at the design points, when a particular SM is used, is:

$$\Delta MSE = MSE_S - MSE_E = \sigma^2 (p - p_1)(R_C - 1) \tag{10}$$

The model with the smallest value of ΔMSE will provide the best predictions, on average, at the design points. Due to the $(p - p_1)$ term in Eqn. (10), when two SMs contain different numbers of parameters (i.e. different values of p_1), the SM with the lower value of R_C may not correspond to the lower MSE_S. As a result, we propose to use the following corrected critical ratio R_{CC} for comparing several models with different number of parameters:

$$R_{CC} = \frac{(MSE_S - MSE_E)/n}{\sigma^2} = \frac{p - p_1}{n}(R_C - 1) \tag{11}$$

R_{CC} is the increase in MSE (per data point) arising from the selection of a candidate SM (rather than the EM) normalized by the noise variance. The true value of R_{CC} corresponding to the EM is zero, and the model with the lowest value of R_{CC} gives the best predictions at the

design points. When the available data are informative enough to support the use of the EM, R_C for all SMs will tend to be larger than 1, and the corresponding R_{CC} will be positive. In situations when the available data are limited, R_{CC} for some SMs will tend to be negative, indicating that these SMs will give better predictions than the EM. The SM with lowest value of R_{CC} will give the best predictions in terms of MSE.

Based on the relationship between R_C and R_{CC} in Eqn. (11), estimates for R_{CC} can be obtained using $\hat{R}_C$. Unfortunately, $\hat{R}_C$ given in Eqn. (7) or (9) is biased and has a large variance (Kubokawa et al., 1993). Improved point estimates for R_C can be obtained using various estimators for the noncentrality parameter λ (Pandey and Rahman, 1971; Kubokawa et al., 1993). A brief summary is provided in the Appendix.

When σ^2 is unknown, we propose that the following truncated estimator for R_C should be used:

$$\hat{R}_{CK} = max\left(\frac{n-p-2}{n-p}\hat{R}_C - 1, \frac{2(n-p-2)}{(p-p_1+2)(n-p)}\hat{R}_C\right) \quad (12)$$

where the subscript K indicates that this estimator was derived using the improved estimator for λ developed by Kubokawa et al. (1993). Note that $\hat{R}_C$, from Eqn. (7), follows a noncentral F distribution. In situations when σ^2 is known, the appropriate truncated estimator is:

$$\hat{R}_{CK} = max\left(\hat{R}_C - 1, \frac{2}{p-p_1+2}\hat{R}_C\right) \quad (13)$$

where $\hat{R}_C$ is obtained from Eqn. (9) and $(p-p_1)\hat{R}_C$ follows a noncentral χ^2 distribution.

The truncated estimators in Eqns. (12) and (13) have lower MSE than the original estimators in Eqns. (7) and (9), and are less computationally demanding than a corresponding maximum likelihood estimator based on the method of Pandey and Rahman (1971). As a result, we propose that modellers should select the best model using:

$$\hat{R}_{CC} = \frac{p - p_1}{n}\left(\hat{R}_{CK} - 1\right) \tag{14}$$

The candidate model (either an SM or the EM) with the lowest value of $\hat{R}_{CC}$ is expected to give the lowest total mean-squared-prediction-error at the design points. We will now extend this new model-selection criterion for selection of univariate and multivariate nonlinear models, which are of greater interest to chemical engineers than are univariate linear models.

3. Extension to Selection of Univariate Nonlinear Models

In the nonlinear case, the EM has the form

$$y = f(X, \theta) + \varepsilon \tag{15}$$

where $f(X, \theta)$ is nonlinear in some or all of the parameters θ, and ε is independently and identically distributed with zero mean and constant variance σ^2. Unlike the linear case, numerical optimization is required to obtain the parameter estimates $\hat{\theta}$ (Seber and Wild, 2003). When there are too many unknown parameters and the available data are limited (e.g., data may be noisy, or may be obtained from poorly designed experiments), SMs, which contain only a subset of the unknown parameters, are often preferred, so that difficulties associated with poor numerical conditioning can be avoided. These candidate SMs can be formulated, either by making different assumptions in the model formulation process, or by leaving some parameters fixed at their initial guesses. Good initial guesses can be obtained based on the available data (Bates and Watts, 1988), the modellers' engineering knowledge and experience, or from similar studies in the literature.

Using a SM with a small set of parameters can significantly reduce the complexity and nonlinearity of the model, as well as the variability of model predictions. However, use of the SM will introduce bias in model predictions. For the nonlinear model described by Eqn. (15), the MSE, which accounts for bias and variance, can be defined as (Rice, 1995)

$$MSE(\hat{y}) = E\left(\left(\hat{y} - f(X,\theta)\right)^T\left(\hat{y} - f(X,\theta)\right)\right)$$
$$= \left(E(\hat{y}) - f(X,\theta)\right)^T\left(E(\hat{y}) - f(X,\theta)\right) + tr\left(Cov(\hat{y})\right) \quad (16)$$

where $\hat{y} = f\left(X,\hat{\theta}\right)$. For nonlinear models, R_{CC} is defined as

$$R_{CC} = \frac{(MSE_S - MSE_E)/n}{\sigma^2} \quad (17)$$

This expression can be compared to Eqn. (11) for linear models. For a given set of candidate models, the one with lowest value of R_{CC} corresponds to the lowest mean-squared-prediction-error, and therefore, should be selected as the best model. For nonlinear univariate models, R_{CC} can also be written as:

$$R_{CC} = \frac{p - p_1}{n}(R_C - 1) \quad (18)$$

where R_C, based on linearization of the nonlinear model, is approximately the squared bias introduced by estimating only a subset of parameters, divided by the associated variance reduction. Unfortunately, for nonlinear models, no exact explicit expression can be written for R_C. When σ^2 is unknown, R_C can be estimated from the data using the likelihood ratio statistic:

$$\hat{R}_C = \frac{(SSE_S - SSE_E)/(p - p_1)}{SSE_E/(n - p)} \quad (19)$$

which is the same as Eqn. (7) for univariate linear models. When σ^2 is known, R_C can be estimated using the right-hand side of Eqn. (9). As a result, the associated truncated estimator $\hat{R}_{CK}$, which is defined in Eqn. (12) or (13), can also be derived, so that R_{CC} can be estimated as:

$$\hat{R}_{CC} = \frac{p - p_1}{n}\left(\hat{R}_{CK} - 1\right) \quad (20)$$

The candidate nonlinear model with the lowest value of $\hat{R}_{CC}$ is expected to give predictions with the lowest total MSE.

The proposed $\hat{R}_{CC}$ criterion for selection of nonlinear univariate models relies on the assumption that $\hat{R}_C$ from Eqn. (19) follows a noncentral F distribution. Gallant (1987) showed that likelihood ratio statistics for nonlinear models, like the one on the right-hand-side of Eqn. (19), can be adequately described by a noncentral F distributions with noncentrality parameter $\lambda = (p - p_1)R_C$. Calculation of $\hat{R}_C$ using Eqn. (19) requires SSE_E, the sum of squared residuals from the EM. In situations where it is impossible to estimate all of the unknown parameters in the EM due to problems of ill conditioning, the value of SSE_E can be approximated using a SM with a sufficiently large number of parameters, so that estimation of additional parameters does not produce a noticeable improvement in the objective function for parameter estimation.

In the next section, Monte Carlo simulations are performed using the Lubricant model of Witt (1974) described by Bates and Watts (1988) to demonstrate: 1) the validity of the approximate noncentral F distribution for $\hat{R}_C$ in Eqn. (19); 2) the effectiveness of the proposed MSE-based criterion for selecting the best nonlinear univariate model; and 3) the effects of various factors (e.g., noise variance, number of data points, initial parameter guesses) on the selection of the best model.

3.1 Example: Lubricant Model (Witt, 1974; Bates and Watts, 1988)

The Lubricant model predicts the logarithm of the kinematic viscosity of a lubricant as a function of temperature (°C) and pressure (atm/1000). This relationship is described by the following empirical model:

$$y = f(X, \theta) + \varepsilon \tag{21}$$

with

$$f(X,\theta)=\frac{\theta_1}{\theta_2+x_1}+\theta_3x_2+\theta_4x_2^2+\theta_5x_2^3+(\theta_6+\theta_7x_2^2)x_2exp\left(\frac{-x_1}{\theta_8+\theta_9x_2^2}\right) \quad (22)$$

where x_1 is temperature and x_2 is pressure. The additive noise ε is assumed to be independently and identically distributed following a Normal distribution with mean zero and constant variance σ^2 (Linssen, 1975). There are $p=9$ unknown parameters in the EM (Eqn. (22)). The original data set consists of $n=53$ data points obtained at four temperature settings (0℃, 25℃, 37.8℃ and 98.9℃) and a variety of pressure settings ranging from 1 atm to 7469.35 atm (Bates and Watts, 1988).

In the Monte Carlo simulations used for testing the performance of $\hat{R}_{CC}$, the true noise variance is set as

$$\sigma^2=0.002 \quad (23)$$

which we estimated based on the original data set. Note that, this true value of σ^2 is only used for generating the data in Monte Carlo simulations. We assume that σ^2 is unknown when selecting the best model.

The true parameter values θ_{true} and the initial parameter guesses θ^0 were set at the values given in Table 1. These values of θ_{true} were obtained by rounding off parameter estimates from the EM using the original data. The third row of Table 1 shows the difference between θ_{true} and θ^0 as a multiple of the standard error for the parameter estimates. For example, the initial guess of -0.3 for θ_4^0 is 8.00 standard errors away from the true value of -0.2 used to generate the simulated data.

(Table 1)

Set of Candidate Models

To demonstrate the usefulness of the proposed MSE-based criterion, we focus on a set of arbitrarily chosen candidate models, which is shown in Table 2. The check marks "√" indicate that the corresponding parameter is estimated in the candidate model. Parameters without a check mark were held at their initial guesses (second row in Table 1) and were not estimated in the corresponding SMs.

(Table 2)

Assessing the Validity of the Noncentral F Distribution Approximation for $\widehat{R}_C$

The assumption that $\hat{R}_C$ (Eqn. (19)) adequately follows a noncentral F distribution is a key requirement in the development of the proposed $\hat{R}_{CC}$ criterion for selecting nonlinear models. In this section, the validity of this approximation is tested for the nonlinear SMs in Table 2. 10000 Monte Carlo simulations with different additive random noise sequences were used to generate simulated data from the model in Eqns. (21) and (22), using the noise variance in Eqn. (23), the true parameter values in Table 1 and the input settings from the original data set.

For each generated data set, $\hat{R}_C$ was calculated for each candidate model. Empirical and theoretical cumulative distributions of $\hat{R}_C$ for SM_1 are compared in Fig. 1. The empirical distribution was obtained by sorting the 10000 values for $\hat{R}_C$ from lowest to highest and plotting the fraction of the 10000 values that is below each possible value of $\hat{R}_C$. The theoretical distribution function was calculated using the noncentral F distribution function in MATLAB™ with $\lambda = (p - p_1)R_C$. The true value of R_C was calculated using Eqns. (17) and (18), where the MSE was approximated using sample biases and sample variances from the complete set of 10000 predictions.

Fig. 1 shows a close match between the empirical and the theoretical curves. Note that similar results were observed for all of the other SMs in Table 2, confirming that the noncentral F distribution is a good approximation for the distribution of $\hat{R}_C$ obtained using these nonlinear models. This result demonstrates that it is appropriate to use the truncated Kubokawa estimator when computing $\hat{R}_{CC}$, which is similar to the more general conclusion of Gallant (1987).

(Figure 1)

Model-Selection based on Original Data and $\widehat{R}_{CC}$

In this section, $\hat{R}_{CC}$ is used to select the best model based on the original data set, starting from the initial parameter guesses given in Table 2. In the first step, each candidate model was fitted using nonlinear least squares, and the corresponding sum of squared residuals was calculated, and $\hat{R}_{CC}$ values for each model were obtained using Eqns. (12), (19) and (20). Based on the original data set, the EM, with all the nine estimated parameters has the smallest $\hat{R}_{CC}$ value and is therefore selected as the best model. This is not surprising, due to the fact that the modeller used considerable effort and expertise when selecting the model form that we consider as the EM (Witt, 1974; Bates and Watts, 1988).

In the next section, the performance of the proposed $\hat{R}_{CC}$ model-selection criterion is illustrated using Monte-Carlo simulations involving situations with different noise variances, numbers of data points, and initial parameter guesses that make it more difficult to estimate all of the parameters in the Lubricant model.

Performance of the Proposed Model-Selection Criterion

To demonstrate the effectiveness of the proposed $\hat{R}_{CC}$ criterion for selecting the model with

the lowest total mean-squared-prediction-error, four sets of Monte Carlo simulations were performed. In the first set (Case 1), the noise setting is the same as in Eqn. (23) and data obtained at all four temperature settings are used for parameter estimation. In the second set (Case 2), the noise variance was increased by a factor of 5, making it more difficult to obtain good parameter estimates. In the third set (Case 3), the simulation settings were the same as in Case 1, except that simulated data obtained at $T = 98.9℃$ were not available for parameter estimation. In the fourth set (Case 4), the simulation settings were the same as in Case 3, but the initial guess for θ_5 was changed from -0.022 to 0, which is farther away from the true value of -0.02. Since θ_5 is held constant in SM_1, SM_3 and SM_4, by setting $\theta_5 = 0$, the corresponding cubic term ($\theta_5 x_2^3$) in the model is deleted. As a result, these SMs have a simpler model structure than the EM.

In each case, Monte Carlo simulations were performed 10000 times using different random noise sequences. Sample means and sample variances from the 10000 sets of predictions were used to compute theoretical values of MSE and R_{CC} (Eqn. (17)), which are shown in Table 3 for each candidate model in all of the four cases considered. The smallest MSE and R_{CC} values, which correspond to the best model, are highlighted in bold. The value of R_{CC} for the EM is zero by definition. The results in Table 3 indicate that the EM will give the best model predictions, on average, using the settings from Cases 1 and 4, and that SM_4, which has two fewer parameters, is preferred in Cases 2 and 3, when the data are less informative. Simplified models SM_1, SM_2 and SM_3 will give worse predictions, on average, than the EM in all four Cases, as indicated by the larger values of MSE and positive values of R_{CC} in Table 3. (Table 3)

Table 4 shows the fraction of the time that each model was selected using the 10000 simulated data sets. Results for the models that were selected most often are highlighted in

bold. In Case 1, the EM, which is theoretically the best model according to Table 3, was selected as the best model 71.14% of the time and SM_4 was selected 28.86% of the time. In Cases 2 and 3, the best model SM_4 was selected most often using $\hat{R}_{CC}$ as the selection criterion, and in Case 4, the EM was selected most often due to the poor initial guess for θ_5. These results demonstrate the effectiveness of the proposed model selection criterion, even in situations when the difference in MSE between the best model and the second-best model is small (Case 3). These simulation results also confirm that when data are noisy (Case 2) or few data points are available (Case 3), a simpler model tends to give better predictions. When the initial guess for a particular parameter is poor, as in Case 4, the proposed selection criterion tends to automatically select a model wherein that parameter is estimated.

For comparison, Table 4 also shows the frequencies for each candidate model being selected using the Bayesian Information Criterion (BIC). In a previous article, we compared the tendencies of nine different model-selection criteria for selecting SMs and determined that BIC did a reliable job of selecting the best model, with the lowest mean-squared-prediction-error, for the particular example studied (Wu et al., 2009). BIC was computed for each SM from (McQuarrie and Tsai, 1998):

$$BIC = log\left(\frac{SSE_S}{n}\right) + \frac{log(n)}{n}p_1 \tag{24}$$

Use of BIC does not require assumptions about the true model structure, which is an advantage over $\hat{R}_{CC}$ in situations where an EM is not available. The performance of BIC and $\hat{R}_{CC}$ for selecting the nonlinear univariate model with lowest mean-squared-prediction-error is compared in the discussion below.

(Table 4)

Comparison of the $\hat{R}_{CC}$ and BIC results indicate that, in the particular example studied, the BIC tends to prefer simpler models. The BIC selected SM_4 65.63% of the time in Case 1 and

selected the more complex EM preferred by $\hat{R}_{CC}$ only 34.37% of the time. In Case 3, the BIC selected SM_1, with only 4 parameters, most often, whereas $\hat{R}_{CC}$ selected the more complex SM_4 most often. In all four Cases shown in Table 4, the $\hat{R}_{CC}$ criterion selected the best model (with the lowest theoretical expected total MSE for model predictions) most often, whereas the BIC only selected the best model in Cases 2 and 4.

Fig. 2 shows the sample means and 95% empirical confidence intervals for the per cent error in the model predictions obtained at each experimental setting. These results were obtained using the 10000 Monte Carlo simulations for Case 3 and compares the quality of the predictions obtained from SM_1, SM_4 and the EM. Note that SM_1 was selected most often as the best model using the BIC, whereas SM_4 was selected most often using the proposed criterion $\hat{R}_{CC}$. As expected, the EM, which was used to generate the data, gives unbiased model predictions (i.e., the sample mean is approximately zero at all design points). Also, as expected, the predictions obtained from the EM have wider 95% confidence intervals, corresponding to larger prediction variances. Predictions obtained using SM_4, which is the best model in terms of total MSE for the model predictions has narrower confidence intervals than the EM and has relatively small bias. Predictions from SM_1, which was preferred by BIC, have even smaller variance than predictions from SM_4, but larger bias, on average. Data settings that result in particularly biased predictions from SM_1 are the settings for experiments 20, 21, 30 and 35. The bias introduced by using SM_4 is considerably smaller at all experimental settings, confirming the effectiveness of the proposed $\hat{R}_{CC}$ criterion for this example problem.

(Figure 2)

4. Extension to Selection of Multivariate Models

Many models that appear in the engineering literature have multiple types of response variables (e.g., temperatures, pressures, concentrations, yields). If the model form is complicated, or the data available are not sufficiently informative, estimating all of the unknown parameters may be very difficult or even impossible (e.g., Kou et al., 2005a, b; Ben Zvi et al., 2004). In these complicated nonlinear situations, there are often many competing simplified models that could be used, depending on the simplifying assumptions that are made and the subset of parameters that is estimated (Chu and Hahn, 2008; Lund and Foss, 2008; Thompson et al., 2007, 2009). The difference in complexity and nonlinearity between candidate SMs and the corresponding EM can be substantial. For example, Kou et al. (2005a) developed an EM for ethylene-hexene copolymerization that had 55 parameters, and chose a SM with only 37 parameters because of the limited data available for parameter estimation. Note that Kou et al. (2005a) had many difficulties deciding how many parameters to estimate and how many to hold constant at their initial values.

We now extend the proposed MSE-based model-selection criterion to the selection of simplified multivariate nonlinear models.

Assume that a model of the form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{bmatrix} = \begin{bmatrix} f_1(X,\theta) \\ f_2(X,\theta) \\ \vdots \\ f_d(X,\theta) \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_d \end{bmatrix} \tag{25}$$

can describe the behaviour of the process, where there are d different response variables. Eqn. (25) may contain a set of algebraic equations or may be the numerical solution of a set of differential and algebraic equations.

Responses obtained using n different sets of experimental conditions can be stacked vertically in a "rolled-out" format (Seber and Wild, 2003), so that n responses for the first variable are at the top, followed by n responses for the second variable, and so on, to give

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n} \\ y_{21} \\ \vdots \\ y_{2n} \\ \vdots \\ y_{dn} \end{bmatrix} = \begin{bmatrix} f_1(X_1, \theta) \\ f_1(X_2, \theta) \\ \vdots \\ f_1(X_n, \theta) \\ f_2(X_1, \theta) \\ \vdots \\ f_2(X_n, \theta) \\ \vdots \\ f_d(X_n, \theta) \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1n} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n} \\ \vdots \\ \varepsilon_{dn} \end{bmatrix} \tag{26}$$

It is convenient to express these equations in the form

$$\boldsymbol{y} = \boldsymbol{f}(X, \theta) + \boldsymbol{\varepsilon} \tag{27}$$

where there are $N = nd$ elements in $\boldsymbol{y}$ if all measurements are available for each set of independent variables. If there are missing values for some measurements at some settings, then N is the total number of data values available for parameter estimation.

Due to the limited information content in the available data, we focus on situations where the noise in different response variables is independent, and ε_i $(i = 1,2, \dots, d)$ is independently and identically distributed following a Normal distribution with zero mean and known variance σ_i^2. Prior information about σ_i^2 may have been obtained from repeated experiments on a similar system. Situations when σ_i^2 is unknown are discussed at the end of this section.

Based on the above assumptions, we have

$$\boldsymbol{\varepsilon} \sim N(0, V) \tag{28}$$

with variance-covariance matrix

$$V = \begin{bmatrix} \sigma_1^2 I_n & 0 & \cdots & 0 \\ 0 & \sigma_2^2 I_n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 I_n \end{bmatrix} \tag{29}$$

where I_n is an $n \times n$ identity matrix. The use of V with diagonal structure in Eqn. (30) is for illustration purposes, and the following analysis is also applicable in situations where the noise in different response variables is correlated and the variance-covariance matrix is known.

To use the proposed criterion for selecting multivariate simplified models, the model in Eqn. (29) is scaled as:

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{f}}(X, \theta) + \tilde{\boldsymbol{\varepsilon}} \tag{30}$$

where

$$\tilde{\boldsymbol{y}} = L\boldsymbol{y} \quad \tilde{\boldsymbol{f}}(X, \theta) = L\boldsymbol{f}(X, \theta) \quad \tilde{\boldsymbol{\varepsilon}} = L\boldsymbol{\varepsilon} \tag{31}$$

The scaling matrix is $L = (U^T)^{-1}$ where U is an upper triangular matrix obtained from Cholesky decomposition of the variance-covariance matrix V where $U^T U = V$.

In the selection of multivariate models, we focus on selecting the best model with lowest MSE for the scaled predictions, $\hat{\tilde{\boldsymbol{y}}} = \tilde{\boldsymbol{f}}\left(X, \hat{\theta}\right)$, where $\hat{\theta}$ is obtained by fitting the scaled model in Eqn. (30) using nonlinear least squares. Note that, after putting the multivariate model in "rolled-out" format and scaling using known variances, we formulate a univariate nonlinear model, where the noise, $\tilde{\boldsymbol{\varepsilon}}$, is independently and identically distributed following a standard Normal distribution. Therefore, the results derived in the previous section for selecting nonlinear univariate models can be used directly for selecting multivariate models. Note that the parameter estimates obtained using ordinary least squares and the scaled model are identical to those that would be obtained using generalized least squares based on the original un-scaled multivariate model (Seber and Wild, 2003).

Because the variance of the noise is known (i.e., $\sigma^2 = 1$ after scaling), R_C in Eqn. (18) can be estimated using:

$$\hat{R}_C = (SSE_S - SSE_E)/(p - p_1) \tag{32}$$

where "SSE" is the sum of squared residuals for the scaled model. Note that Eqn. (32) is the same as the expression in Eqn. (9) for univariate linear models with $\sigma^2 = 1$. Similar to univariate models, the distribution of $(p - p_1)\hat{R}_C$ can be approximated by a noncentral χ^2 distribution, with $(p - p_1)$ degrees of freedom and noncentrality parameter $\lambda = (p - p_1)R_C$. As a result, the truncated estimator given in Eqn. (13) can be used, and R_{CC} in Eqn. (17) can be estimated as

$$\hat{R}_{CC} = \frac{p - p_1}{N}\left(\hat{R}_{CK} - 1\right) \tag{33}$$

where N is the total number of data values available for parameter estimation. The candidate model with the lowest value of $\hat{R}_{CC}$ is expected to give the lowest total MSE for the scaled predictions.

In next section, a dynamic α-pinene model (Fuguitt and Hawkins, 1947; Box et al., 1973) is used in Monte Carlo simulations to demonstrate: 1) the quality of the approximate noncentral χ^2 distribution for $(p - p_1)\hat{R}_C$; and 2) the performance of the proposed MSE-based criterion for selecting simplified multivariate nonlinear models.

4.1 Example: α-Pinene Model

The α-pinene thermal isomerization process studied by Fuguitt and Hawkins (1947) has three measured responses and the EM consists of five ordinary differential equations (ODEs) (Bates and Watts, 1988):

$$\begin{aligned}
\frac{df_1}{dt} &= -(\theta_1+\theta_2)f_1 && y_1 = f_1+\varepsilon_1\\
\frac{df_2}{dt} &= \theta_1 f_1 &&\\
\frac{df_3}{dt} &= \theta_2 f_1 - (\theta_3+\theta_4)f_3 + \theta_5 f_5 && y_3 = f_3+\varepsilon_3\\
\frac{df_4}{dt} &= \theta_3 f_3 &&\\
\frac{df_5}{dt} &= \theta_4 f_3 - \theta_5 f_5 && y_5 = f_5+\varepsilon_5
\end{aligned} \tag{34}$$

where f_i $(i = 1, 2, \dots, 5)$ correspond to the concentrations of α-pinene, dipentene, alloocimene, pyronene, and dimer, respectively, (in mole %) taken at various times. Only three independent measurements are available (y_1, y_3 and y_5). Although the right-hand sides of the ODEs are linear in the parameters, the predicted responses are nonlinear in the parameters due to exponentials that appear in the analytical solution for the model equations (Box et al., 1973).

The initial values used in the experimental runs are

$$\begin{aligned}
f_1^0 &= 100\%\\
f_2^0 = f_3^0 = f_4^0 = f_5^0 &= 0\%
\end{aligned} \tag{35}$$

In the simulated experiments used to test the proposed MSE-based MSC, the times (in minutes) at which simulated measurements were generated match the times used by Fuguitt and Hawkins:

$$t = 10 \times (123 \quad 306 \quad 492 \quad 780 \quad 1068 \quad 1503 \quad 2262 \quad 3642)^T \tag{36}$$

When generating the simulated experiments, it was assumed that the additive noise in y_1, y_3 and y_5 is Normally distributed with zero mean, and variance-covariance matrix:

$$V = Cov\left(\begin{bmatrix}\varepsilon_1\\ \varepsilon_3\\ \varepsilon_5\end{bmatrix}\right) = \begin{bmatrix}0.6I_n & 0 & 0\\ 0 & 0.3I_n & 0\\ 0 & 0 & 0.8I_n\end{bmatrix} \tag{37}$$

These variances are consistent with the original data. In the analysis below, V is assumed to be known both for model scaling and for model selection.

In this example, there are $d = 3$ response variables and $n = 8$ observations for each response, so there are $N = nd = 24$ data points available for parameter estimation.

The true parameter values θ_{true} used to conduct the Monte Carlo simulations were obtained by fitting the model to the original data and rounding the resulting parameter values. These true values are listed in Table 5, along with initial parameter guesses used for parameter estimation from the 10000 simulated data sets. The last row of Table 5 shows the deviation between θ_{true} and θ^0 as a multiple of the standard error for the parameter estimates, which was computed from the 10000 simulations.

(Table 5)

To demonstrate the usefulness of the proposed MSE-based criterion, a set of candidate models was arbitrarily chosen by fixing some parameters and estimating the others, as shown in Table 6.

(Table 6)

Assessing the Validity of the Noncentral χ^2 Distribution Approximation for $(p - p_1)\hat{R}_C$

The validity of the noncentral χ^2 approximation for $(p - p_1)\hat{R}_C$ determines the effectiveness of the Kubokawa estimator for R_C in Eqn. (13). 10000 Monte Carlo simulations with different additive random noise sequences were used to generate simulated data based on the above settings. For each generated data set, $\hat{R}_C$ was calculated for each SM using Eqn. (33). The noncentrality parameter $\lambda = (p - p_1)R_C$ for the theoretical non-central χ^2 distributions was calculated from the set of 10000 simulations using the sample mean and sample variance for the scaled predictions.

Fig. 3 shows the empirical cumulative distributions of $(p - p_1)\hat{R}_C$ for SM_1 along with the theoretical χ^2 distribution. The empirical curve matches the theoretical curve closely,

confirming that the noncentral χ^2 distribution is a good approximation for the distribution of $(p - p_1)\hat{R}_C$ obtained using this nonlinear model. Similar results were observed for the other SMs in Table 6, as would be expected based on the results of Gallant (1987).

(Figure 3)

Performance of the Proposed Model-Selection Criterion

We now examine the performance of the proposed $\hat{R}_{CC}$ criterion for selecting the best model from the set of candidate models in Table 6. Table 7 shows the theoretical value of the total MSE and R_{CC} for each candidate model, computed using Eqns. (16) and (17). Note that the EM, which has the smallest value of MSE and R_{CC}, is the best model in terms of mean-squared-prediction-error. Based on 10000 Monte Carlo simulations, frequencies of each model being selected using $\hat{R}_{CC}$ are provided in the third row of Table 7. The proposed criterion selects the EM more often than it selects the other models (42.35% of the time), which demonstrates the effectiveness of this criterion for determining the best model.

Table 7 also shows the frequencies of each model being selected using BIC. These results were obtained using Eqn. (25) based on the scaled model in Eqn. (31). Note that, BIC selected SM_2 most often (54.25% of the time), and selected the best model (the EM) only 8.16% of the time.

(Table 7)

Fig. 4 shows the sample means and 95% empirical confidence intervals for the per cent error in the scaled model predictions obtained at each experimental setting. These results were generated to compare the quality of the predictions obtained from SM_2 and the EM. Note that SM_2 was selected most often as the best model by the BIC, whereas the EM was preferred using the proposed criterion $\hat{R}_{CC}$. As expected, the EM gives unbiased model

predictions. Predictions obtained using SM_2 have narrower confidence intervals than the EM, indicating smaller prediction variances, and have considerable bias for some data settings, especially predictions for data settings from 15 to 19. These results confirm the tendency of the BIC to select simpler models, with higher mean-squared-error, than are selected by the proposed $\hat{R}_{CC}$ criterion.

(Figure 4)

4.2 Selecting Multivariate Models when Noise Variances are Unknown

The analysis in the previous section was based on the assumption that the variance-covariance matrix for the model errors was known *a priori* by the modeller. Although this assumption is valid in situations where there has been considerable prior experimentation on similar systems, there are many situations where the modeller will have limited information about the variances of measured responses. There are both practical and theoretical challenges to extending the proposed $\hat{R}_{CC}$ criterion to selection of models when the variance-covariance matrix is unknown, especially in the situations where the information content in the available data is too weak to support reliable estimation of all of the unknown parameters. In these less-than-ideal situations, attempting to estimate the parameters and the noise variances using, e.g., iteratively reweighted least squares or maximum likelihood methods, is not feasible (Seber and Wild, 2003).

A Bayesian solution for dealing with uncertainty about variances of response variables would be to specify prior distributions for uncertain elements in the variance-covariance matrix, incorporating any knowledge about uncertainty that might be available to the modeller. Random sampling from the prior distributions could be used to scale the model, estimate the parameters and then calculate many different values of $\hat{R}_{CC}$ for the candidate models under

consideration. After a large number of re-sampling, parameter estimation and model selection calculations, the final best model could be determined as the one that was selected most often. This brute-force approach may not be computationally feasible for complex models of chemical processes, where computation times would be prohibitive if the time required to solve the model equations and to estimate the parameters is considerable. This methodology will be tested for relatively simple models in our future work.

5. Conclusions

Parameter estimation in complex models of chemical processes can be difficult, especially when there are many unknown parameters and the available data for parameter estimation are limited (e.g., when noisy data are obtained from poorly designed experiments). In these situations, simplified models with a reduced set of parameters to estimate are often preferred to complex models because they can give more reliable predictions. Candidate simplified models can be obtained by making different assumptions during model formulation or by fixing some parameters at nominal values. Modellers want to determine which simplified model will result in the best predictions, given the available data for parameter estimation.

In this article, a reliable and easy-to-use model-selection criterion is developed to assist modellers in the selection of simplified linear or nonlinear models. The new criterion $\hat{R}_{CC}$ is derived by using total mean-squared-error (MSE) to account for bias and variance in the model predictions. Calculation of $\hat{R}_{CC}$ requires the modeller to have knowledge of (or to make an assumption about) the structure of a full model that is capable of describing the underlying behaviour of the process. The effectiveness of this new criterion is demonstrated theoretically and using Monte Carlo simulations involving nonlinear single-response and multi-response models.

The performance of the $\hat{R}_{CC}$ criterion is compared with that of the BIC. Both criteria are effective, in that they tend to select simplified models, rather than complex models when data are limited. For the examples studied, the $\hat{R}_{CC}$ criterion consistently selects simplified models that give the total lowest mean-squared-prediction-errors. In some situations, the BIC tends to select overly simplified models rather than the best model.

The proposed $\hat{R}_{CC}$ criterion can be applied to the selection of multi-response models when variances for the different response variables are assumed to be known. Difficulties associated with multivariate model selection with unknown noise variances are discussed, and possible approaches are suggested and will be tested in our future work.

Appendix: Summary of Various Estimators for the Noncentrality Parameter

Improved estimators for R_C can be derived based on various estimators for the noncentrality parameter λ, which appears in noncentral F and χ^2 distributions. Results from Kubokawa et al. (1993) are summarized below.

Given a random variable S, such that $S \sim \chi_v^2(\lambda)$, the unbiased estimators for λ is

$$\hat{\lambda}_0 = S - v \tag{A1}$$

which can take negative values. The following improved estimator, which results in smaller mean-squared-error and no negative values, was proposed by Kubokawa et al.

$$\hat{\lambda}_K = max\left(\hat{\lambda}_0, \frac{2}{v+2}S\right) \tag{A2}$$

Similarly, in the case when random variable S follows a noncentral F distribution, $S \sim F_{v_1,v_2}(\lambda)$, the uniformly minimum-variance unbiased estimators for λ is

$$\hat{\lambda}_0 = \frac{v_1(v_2-2)}{v_2}S - v_1 \tag{A3}$$

Kubokawa et al. proposed the following truncated estimator, which cannot take negative values and which results in lower mean-squared-error

$$\hat{\lambda}_K = max\left(\hat{\lambda}_0, \frac{2v_1(v_2-2)}{v_2(v_1+2)}\right) \quad \text{(A4)}$$

The truncated estimator for R_C in Eqns. (12) and (13) was derived from Eqns. (A2) and (A4).

Nomenclature

d	number of response variable
e	stochastic component with any model mismatch
f	nonlinear model, deterministic response
log	natural logarithm
n	number of data points for a single response variable
p	total number of unknown parameters
s^2	noise variance estimates
t	time
v	degree of freedom
x	input variable
y	vector of response variable
E	expected value
I_n	$(n \times n)$ identity matrix
L	scaling matrix
N	total number of data points
P	projection matrix
R_C	critical ratio
R_{CC}	corrected critical ratio
S	random variable
U	upper triangular matrix
V	variance-covariance matrix
X	matrix of regression variables
Y	response variable

Greek Symbols

β, θ	unknown parameters
ϵ	stochastic component
λ	noncentrality parameter
σ^2	noise variance

Superscripts

$\widehat{\ }$	estimated value
^{0}	initial values
^{T}	matrix transcript
$^{-1}$	matrix inverse

Subscripts

$_1$	first partitioned part
$_2$	second partitioned part
$_i$	index
$_{true}$	noise-free response or true values
$_E$	extended model
$_K$	Kubokawa estimate
$_S$	simplified model

Abbreviations

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
Cov	Variance-Covariance Matrix
EM	Extended Model
FPE	Final Prediction Error
max	maximum
MSC	Model-Selection Criteria
MSE	Mean-Squared-Error
SM	Simplified Model
tr	trace

Others

$\mathbb{R}^n$	column vector of length n taking real values
$\mathbb{R}^{n\times p}$	$(n \times m)$ matrix taking real values

Acknowledgements

The authors would like to thank Cybernetica, DuPont, Hatch, Matrikon, SAS, MITACS (Mathematics of Information Technology and Complex Systems) and NSERC (TJH) for financial support of this research.

References

Akaike, H. “Information Theory and an Extension of the Maximum Likelihood Principle,” 2. Int. S. Inf. Theor., Ed. PETROV, B. N. and CSAKI F., pp. 267-281. Budapest: Akademia Kiado (1973).

Aris, R. “Mathematical Modeling: A Chemical Engineer’s Perspective,” Academic Press, NY (1999).

Bagajewicz, M. J. and E. Cabrera, “Data Reconciliation in Gas Pipeline Systems,” Ind. Eng. Chem. Res. **42(22)**, 5596-5606 (2003).

Bates, D. M. and D. G. Watts, “Nonlinear Regression Analysis and Its Applications,” John Wiley & Sons, NY (1988).

Beck, J. V. and K. J. Arnold, “Parameter Estimation in Engineering and Science,” John Wiley & Sons, NY (1977).

Ben-Zvi, A., K. McAuley and J. McLellan, “Identifiability Study of a Liquid-Liquid Phase-Transfer Catalyzed Reaction System,” AIChE J., **50(10)**, 2493-2501 (2004).

Box, G. E. P., W. G. Hunter, J. F. MacGregor and J. Erjavec, “Some Problems Associated with the Analysis of Multiresponse data,” Technometrics, **15**, 33-5 (1973).

Burnham, K. P. and D. R. Anderson, “Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach,” 2nd Edn., Springer, NY (2002).

Chang, S., T. D. Waite and A. G. Fane, “A Simplified Model for Trace Organics Removal by Continuous Flow PAC Adsorption/Submerged Membrane Processes,” J. Membrane Sci. **253(1-2)**, 81-87 (2005).

Chu Y. and J. Hahn, “Integrating Parameter Selection with Experimental Design Under Uncertainty for Nonlinear Dynamic Systems,” AIChE J., **54(9)**, 2310-2320 (2008).

Fuguitt, R. E. and J. E. Hawkins, “Rate of Thermal Isomerization of α-pinene in the Liquid Phase,” J. Am. Chem. Soc., **69**, 319-322 (1947).

Gallant A. R., “Nonlinear Statistical Models,” John Wiley & Sons, NY (1987).

Hocking, R. R. “Analysis and Selection of Variables in Linear Regression,” Biometrics **32(1)**, 1-49 (1976).

Konishi, S. and G. Kitagawa, “Information Criteria and Statistical Modeling,” Springer, NY (2008).

Kou B., K. B. McAuley, J. C. C. Hsu and D. W. Bacon, “Mathematical Model and Parameter Estimation for Gas-Phase Ethylene/Hexene Copolymerization with Metallocene Catalyst,” Macromol. Mater. Eng., **290**, 537-557 (2005a).

Kou B., K. B. McAuley, C. C. Hsu, D. W. Bacon and K. Z. Yao, “Mathematical Model and Parameter Estimation for Gas-Phase Ethylene Homopolymerization with Supported Metallocene Catalyst,” Ind. Eng. Chem. Res. **44**, 2428-2442 (2005b).

Kubokawa, T., C. P. Robert and A. K. Md. E. Saleh, “Estimation of Noncentrality Parameters,” Can. J. Stat., **21(1)**, 45-57 (1993).

Linhart, H. and W. Zuccini, “Model Selection,” John Wiley & Sons, NY (1986).

Linssen, H. N., “Nonlinearity Measures: A Case Study,” Stat. Neerl., **29**, 93-99 (1975).

Lund B. F. and B. A. Foss, “Parameter Ranking by Orthogonalization – Applied to Nonlinear Mechanistic Models,” Automatica, **44**, 278-281 (2008).

Lv, P., J. Chang, T. Wang, C. Wu and N. Tsubaki, “A Kinetic Study on Biomass Fast Catalytic Pyrolysis,” Energ. Fuel. **18(6)**, 1865-1869 (2004).

Maria, G. “A Review of Algorithms and Trends in Kinetic Model Identification for Chemical and Biochemical Systems,” Chem. Biochem. Eng. Q. **18(3)**, 195-222 (2004).

Maria, G. “Application of Lumping Analysis in Modeling the Living Systems - a Trade-off between Simplicity and Model Quality,” Chem. Biochem. Eng. Q. **20(4)**, 353-373 (2006).

Mchaweh, A., A. Alsaygh, K. Nasrifar and M. Moshfeghian, “A Simplified Method for Calculating Saturated Liquid Densities,” Fluid Phase Equilib. **224(2)**, 157-167 (2004).

McQuarrie, A. D. R. and C. L. Tsai, “Regression and Time Series Model Selection,” World Scientific, Singapore (1998).

Montgomery, D. C., E. A. Peck and G. G. Vining, “Introduction to Linear Regression Analysis,” 3rd ed, John Wiley & Sons, NY (2001).

Pandey, J. N. and M. Rahman, “The Maximum Likelihood Estimate of the Noncentrality Parameter of a Noncentral F Variate,” Siam J. Math. Anal., **2(2)**, 269-276 (1971).

Perregaard, J., “Model Simplification and Reduction for Simulation and Optimization of Chemical Processes,”. Comput. Chem. Eng. **17(5-6)**, 465-483 (1993).

Rao, C. R. and Y. Wu, “On Model Selection (with Discussion),” in “Model Selection,” P. Lahiri, IMS Lecture Notes – Monograph Series **38**, 1-64 (2001).

Rao, P. “Some Notes on Misspecification in Multiple Regressions,” Am. Stat. **25(5)**, 37-39 (1971).

Rice, J. A., “Mathematical Statistics and Data Analysis,” 2nd Ed., Duxbury Press, Belmont, CA (1995).

Romdhane, M. and C. Tizaoui, “The Kinetic Modeling of a Steam Distillation Unit for the Extraction of Aniseed (Pimpinelia Anisum) Essential Oil,” J. Chem. Technol. Biot. **80(7)**, 759-766 (2005).

Seber, G. A. F. and C. J. Wild, “Nonlinear Regression,” John Wiley & Sons, NY (2003).

Thompson D. E., K. B. McAuley and P. J. McLellan, “A Simplified Model for Prediction of Molecular Weight Distributions in Ethylene-Hexene Copolymerization Using Ziegler-Natta Catalysts,” Macromol. React. Eng. **1**, 523-536 (2007).

Thompson D. E., K. B. McAuley and P. J. McLellan, “Parameter Estimation in a Simplified MWD Model for HDPE Produced by a Ziegler-Natta Catalyst,” Macromol. React. Eng. **3**, 160-177 (2009).

Vitt, L. Die Berechnung physikalischer und thermodynamischer Kennworte von Druckflussigkeiten, sowie die Bestimmung des Gesamtwirkungsgrades an Rumpen under Berucksichtigung der Thermodynamik fur die Druckflussigkei. Ph.D. Thesis, Technological University Eindhoven. 1974.

Wang, F. Y., Z. H. Zhu, P. Massarotto and V. Rudolph, “A Simplified Dynamic Model for Accelerated Methane Residual Recovery from Coals,” Chem. Eng. Sci. **62(12)**, 3268-3275 (2007).

Wang, S. G. and S. C. Chow, “Advanced Linear Models: Theory and Applications,” Marcel Dekker, NY (1994).

Wu, S., K. B. McAuley and T. J. Harris, “The Use of Simplified or Misspecified Models: Linear Case,” Can. J. Chem. Eng. **85**, 386-398 (2007).

Wu, S., K. B. McAuley and T. J. Harris, “Selection of Simplified Models: I. Analysis of Model Selection Criteria Using Mean-Squared Error,” Submitted to Can. J. Chem. Eng. (2009).

Yoshida, H., Y. Takahashi and M. Terashima, “A Simplified Reaction Model for Production of Oil, Amino Acids, and Organic Acids from Fish Meat by Hydrolysis under Sub-Critical and Supercritical Conditions,” J. Chem. Eng. JPN. **36(4)**, 441-448 (2003).

Zhang, P. “Comment on ‘An Asymptotic Theory for Linear Model Selection’”. Stat. Sinica, **7**, 254-258 (1997).

Table 1: True parameter values and initial parameter guesses used in Monte Carlo simulations

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9
$\theta_{i,true}$	1000	200	1.36	-0.2	-0.02	0.4	0.1	50	-0.45
θ_i^0	960	210	1.42	-0.3	-0.022	0.35	0.05	48	-0.49
deviation	1.66	1.88	1.71	8.00	1.49	1.53	34.12	2.41	1.80

Table 2: Candidate Models

Candidate Model	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	Number of Parameters p_1
SM_1		√		√			√	√		4
SM_2			√		√	√	√	√		5
SM_3	√	√		√		√	√	√		6
SM_4	√	√	√	√		√	√		√	7
EM	√	√	√	√	√	√	√	√	√	$p = 9$

Table 3: MSE for Model Predictions and Corresponding True Values of R_{CC} for Each Candidate Model in All Four Cases

		SM_1	SM_2	SM_3	SM_4	EM
Case 1	MSE	0.2934	1.1293	0.2147	0.0261	**0.0184**
	R_{CC}	2.5942	10.4799	1.8517	0.0721	**0**
Case 2	MSE	0.3257	1.1690	0.2627	**0.0821**	0.0905
	R_{CC}	0.4439	2.0350	0.3249	**-0.0157**	0
Case 3	MSE	0.0310	0.9557	0.0308	**0.0175**	0.0182
	R_{CC}	0.1684	12.3355	0.1657	**-0.0088**	0
Case 4	MSE	0.1238	0.9557	0.0943	0.0404	**0.0182**
	R_{CC}	1.3892	12.3355	1.0010	0.2928	**0**

Table 4: Fraction of Each Candidate Model Being Selected Using $\widehat{R}_{CC}$ and BIC

		Fraction of Each Model Being Selected				
		SM_1	SM_2	SM_3	SM_4	EM
Case 1	$\widehat{R}_{CC}$	0	0	0	0.2886	**0.7114**
	BIC	0	0	0	**0.6563**	0.3437
Case 2	$\widehat{R}_{CC}$	0.0037	0	0.0055	**0.7088**	0.2820
	BIC	0.0939	0	0.0106	**0.8515**	0.0440
Case 3	$\widehat{R}_{CC}$	0.2007	0	0.0479	**0.5274**	0.2240
	BIC	**0.5777**	0	0.0235	0.3788	0.0200
Case 4	$\widehat{R}_{CC}$	0	0	0	0.0435	**0.9565**
	BIC	0.0034	0	0.0005	0.2012	**0.7949**

Table 5: true parameter values and initial parameter guesses used in Monte Carlo simulations

	θ_1	θ_2	θ_3	θ_4	θ_5
$\theta_{i,true}$ $(\times 10^{-5})$	6	3	2	28	4
θ_i^0 $(\times 10^{-5})$	5.84	2.65	1.63	24.5	5.5
deviation	2.91	7.16	1.87	1.61	2.02

Table 6: Candidate Models. Parameters indicated by √ are included for estimation in the corresponding SM and the remaining parameters are fixed at their initial guesses in Table 5.

Candidate Model	θ_1	θ_2	θ_3	θ_4	θ_5	Number of Parameters p_1
SM_1	√					1
SM_2		√		√		2
SM_3	√		√		√	3
SM_4	√	√	√		√	4
EM	√	√	√	√	√	$p = 5$

Table 7: True Values of MSE and R_{CC} and the Frequencies of Each Model Being Selected using $\widehat{R}_{CC}$ and BIC

		SM_1	SM_2	SM_3	SM_4	EM
Theoretical Value	MSE	77.2407	9.4627	10.1075	7.3132	**5.0643**
	R_{CC}	3.0073	0.1833	0.2101	0.0937	**0**
Frequencies	$\widehat{R}_{CC}$	0	0.2473	0.1397	0.1895	**0.4235**
	BIC	0	**0.5425**	0.2160	0.1599	0.0816

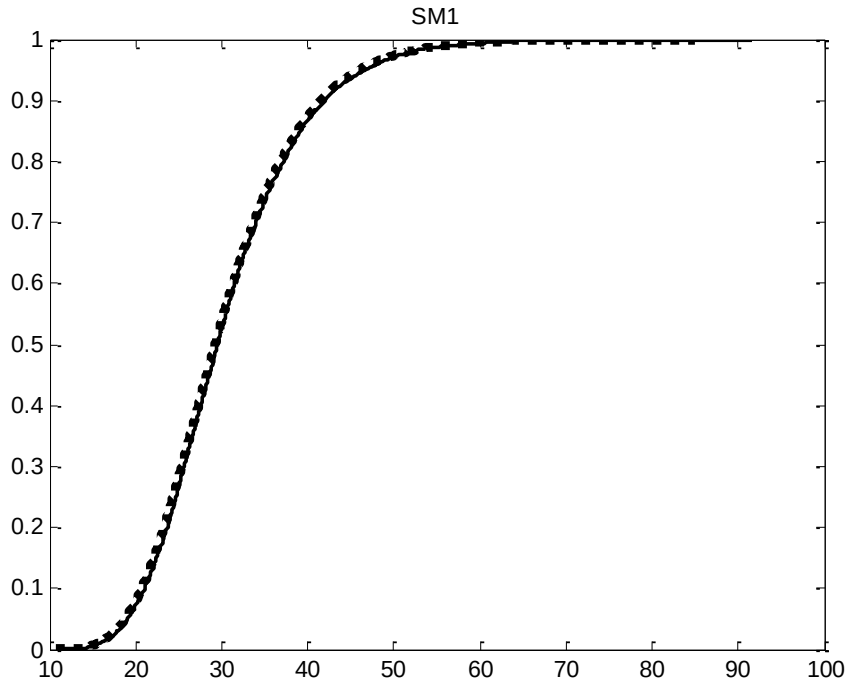

Fig. 1: Comparison of the Theoretical Cumulative Distribution (----) for $\widehat{R}_C$ and the Empirical Distribution (—) Obtained from 10000 Monte Carlo Simulations for SM_1. Note that the two curves are nearly coincident.

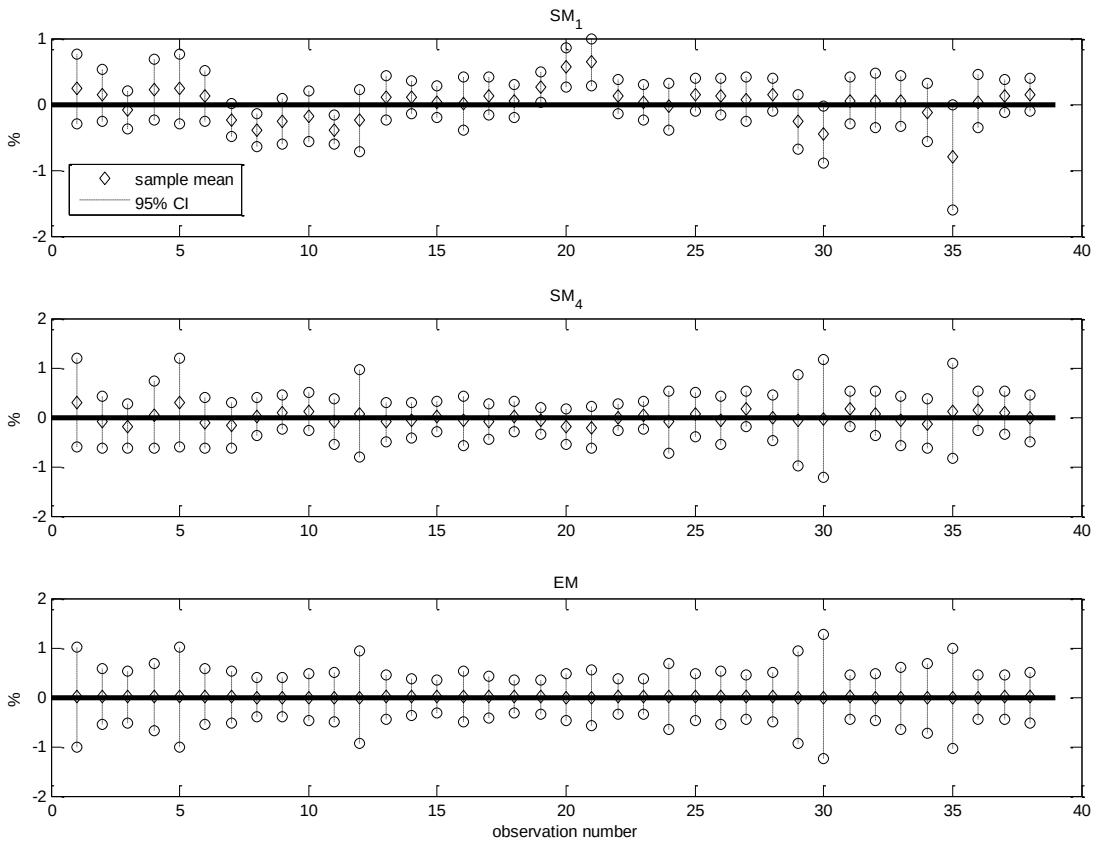

Fig. 2: sample means and 95% empirical confidence intervals (CI) of $\left(f(X,\hat{\theta}) - f(X,\theta)\right)/f(X,\theta)$ for $\mathbf{SM_1}$, $\mathbf{SM_4}$ and the $\mathbf{EM}$ at each prediction point (Case 3)

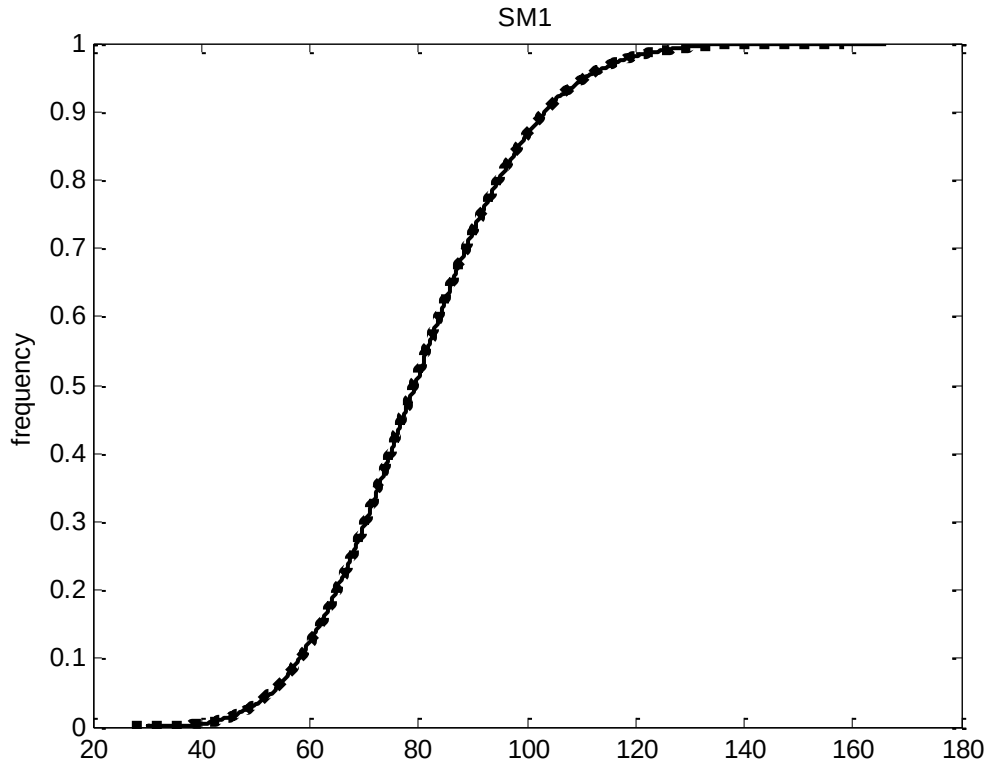

Fig. 3: Comparison of the Theoretical Cumulative Distribution (----) for $\widehat{R}_C$ and the Empirical Distribution (—) Obtained from 10000 Monte Carlo Simulations for SM_1 in Table 4.

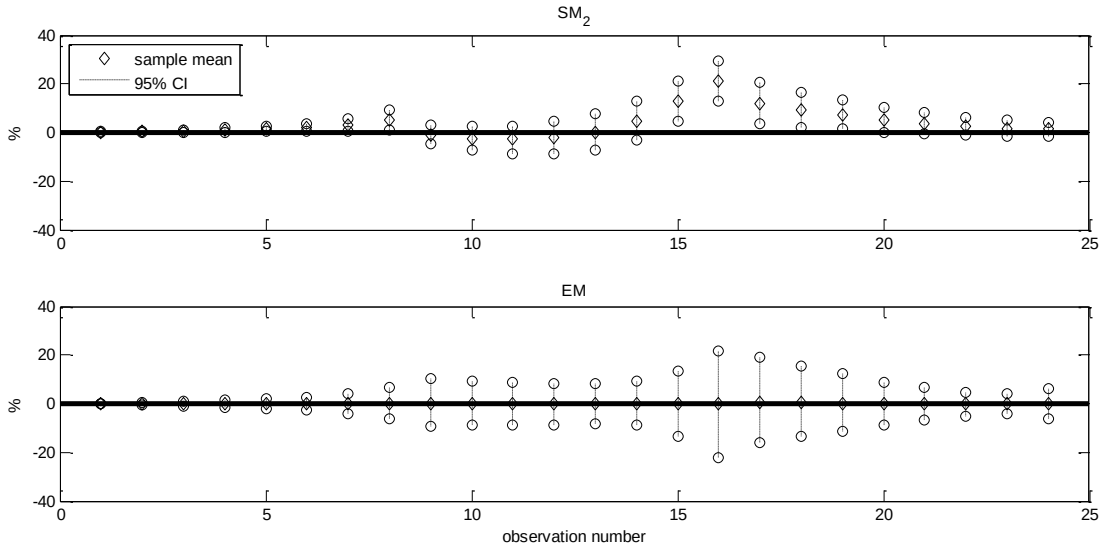

Fig. 4: sample means and 95% empirical confidence intervals (CI) of $\left(\tilde{\boldsymbol{f}}(\boldsymbol{X},\widehat{\boldsymbol{\theta}})-\tilde{\boldsymbol{f}}(\boldsymbol{X},\boldsymbol{\theta})\right)/\tilde{\boldsymbol{f}}(\boldsymbol{X},\boldsymbol{\theta})$ for $\mathbf{SM_1}$ and the EM at each prediction point. Observation numbers 1 to 8 correspond to α-pinene (f_1) predictions, numbers 9 to 16 correspond to alloocimene (f_3) and numbers 17 to 24 correspond to the dimer (f_5).