# IDENTIFICATION OF SEMI-PARAMETRIC HYBRID PROCESS MODELS

Aidong Yang[*,1,2], Elaine Martin[1] and Julian Morris[1]

[1]School of Chemical Engineering and Advanced Materials, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK

[2]Division of Civil, Chemical, and Environmental Engineering, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, GU2 7XH, UK

## Abstract

Hybrid models are mathematical models that comprise both mechanistic and black-box or data-driven components. Typically, the parameters in the mechanistic part of a hybrid model (if any) are assumed to be known. However in this research, a two-level approach is proposed for the identification of hybrid models where some parameters in the mechanistic part of the model are unknown. At the first level, the black-box component is identified using a regularization method with given values for the regularization and mechanistic parameters. At the second level, the regularization and mechanistic parameters are determined simultaneously and optimized according to a specific criterion placed on the predictive performance of the hybrid model. This approach is tested through the modelling of a toluene nitration process, where a support vector machine (SVM) model is used to represent the chemical kinetics, with the mass transfer-related mechanistic parameters being estimated simultaneously. The case study shows that good results can be obtained in terms of both the prediction of the process variables of interest and the estimates of the mechanistic parameters, when the measurement error in the training data is small whilst when the magnitude of the measurement error increases, the accuracy of the estimates of the mechanistic parameters decreases. However, the predictive performance of the resulting hybrid model in the latter case is still acceptable, and can be much better than that attained from the application of a pure black-box model under certain extrapolation conditions.

**Keywords:** Hybrid modelling; Model identification; Semi-parametric models

* Corresponding author. Current address: Faculty of Engineering and Physical Sciences (J2), University of Surrey, Guildford GU2 7XH, UK. Tel.: +44 1483 686577; fax: +44 1483 686581. *E-mail address:* a.yang@surrey.ac.uk (A. Yang).

## 1. Introduction

In recent years, mathematical models for the characterization of chemical processes has become increasingly important for supporting various types of engineering tasks such as process design and process control. Depending on the information available, process modelling may be performed through a number of approaches including mechanistic, black-box (or data-driven), or hybrid modelling. By bringing together both existing mechanistic knowledge and data gathered from the process, a hybrid model that fuses both components has been shown, in a number of applications, to be advantageous when compared with a model formulated from either limited mechanistic knowledge or one constructed solely from the process data (Psichogios, Ungar, 1992; Thompson, Kramer, 1994; Duarte et al., 2004; Oliveira, 2004). The advantages of hybrid models have motivated a number of applications, such as the modelling of batch polymerization reactors (Tian et al., 2001), fermentation processes (Wang et al., 2009; Saraceno et al., 2009) and boilers (Rusinowski, Stanek, 2009). Besides, Teixeira et al. (2007) discussed the general role of hybrid modeling in the combination of systems biology and process engineering.

When modelling a chemical process, the black-box model or the black-box component of a hybrid model will usually have the characteristics of a "universal approximator," i.e. one that is capable of approximating any arbitrary function. Within this group of models, artificial neural networks (e.g. Psichogios, Ungar, 1992; Thompson, Kramer, 1994; Montague, Morris, 1994) have most often been considered. More generally, a black-box model will belong to the family of non-parametric models (Eubank, 1988; Hastie, Tibshirani, 1990; Green, Silverman, 1994). When identifying a non-parametric model, regularization is often applied to address the issue of over-fitting, i.e. a regularization parameter is utilized as a weighting factor for the penalty term in the criterion for training. Thus, the development of a non-parametric model under

regularization typically includes two tasks: the selection of an appropriate value for the *regularization parameter*, and the estimation of the *parameters of the non-parametric model* according to the selected training criterion. A detailed discussion on this topic can be found in the non-parametric regression literature (e.g. Geman et al., 1992; Green, Silverman, 1994).

With respect to the mechanistic component of a hybrid model, the most frequently utilised first principles knowledge are the conservation laws. These materialize in the mass/energy balance equations being incorporated into the model. In contrast, the mechanistic knowledge often missing from a model relates to the so-called constitutive relationships, i.e. those relationships that define the rates of (bio)-chemical reactions and transport phenomena, or those relationships that model the physical properties. In reported hybrid modelling studies (references cited above), black-box models are frequently adopted for approximating the unknown constitutive relationships, whilst the rest of the mechanistic knowledge is assumed to be available. The systematic methods for identifying this type of models can be found in e.g. Kahrs and Marquardt (2008). However, in reality a constitutive relationship may be of a known mechanistic form but has unknown parameters that require to be estimated. This relationship, and those described by non-parametric, block-box models, may require to be combined to give an overarching model. Such a situation has received little (if any) consideration to date in the process modelling literature[†]. Lima et al (2007) reported a framework for establishing semi-mechanistic models by adding empirical elements into a mechanistic model which itself may have unknown parameters. These empirical elements, selected out of an "extension set", tend to be relatively simple, parametric expressions, therefore representing a class of models different from that addressed in this work.

[†]This paper addresses those cases where a black-box model is developed for the modelling of a constitutive relationship. However, the proposed solution approach may also be applicable to other types of hybrid models that involve unknown mechanistic physical parameters.

A hybrid model that combines both a parametric and a non-parametric component can be termed a semi-parametric model. Semi-parametric models have been studied in detail in statistics but primarily in the context of data-driven modelling (cf. Ruppert et al, 2003; Haerdle et al, 2004). In contrast, the semi-parametric models considered in this paper are the result of hybrid modelling, where the model structures are typically derived from mechanistic knowledge and hence fail to conform to the typical semi-parametric model forms that have been studied in statistics, such as the generalized partial linear additive model (Haerdle, et al, 2004).

Investigating regularisation methods for ill-posed problems, Weese and co-workers (Weese, 1993; Roths et al., 2001) studied a type of model identification problems, where both an unknown function $f$ and a number of unknown parameters $a_i$ were to be estimated. More specifically, the model to be identified assumes the following form:

$$g(t) = K(f)(t) + \sum_i a_i h_i(t) \ , \qquad (1)$$

where $K$ is a nonlinear operator of $f$, $h_i$ is a function, $t$ is time. Both $K$ and $h_i$ are known from theory. Regularisation was introduced in the identification process; the regularisation parameter was determined by means of optimisation while the parameters of the finite-dimensional approximator of $f$ as well as the "mechanistic" parameters $a_i$ were identified simultaneously.

The focus of this paper is the identification of semi-parametric hybrid models of chemical processes which are generally different from those represented by Eq. (1). The problem is formulated in Section 2 whilst in Sections 3 and 4, a two-level identification approach and its implementation are described, respectively. This approach is different from the one by Weese and co-workers and allows to incorporate established black-box modelling algorithms into an optimisation framework without changes to these algorithms. An application of the proposed approach is reported in Sections 5 and 6 with conclusions being presented in Section 7.

## 2. Problem Formulation

Consider the following functional relationship as a model, or part of a model, that characterizes a chemical process system:

$$y = f(x, v, p), \tag{2}$$

$$v = h(x), \tag{3}$$

where $x$ and $y$ are vectors of the independent and dependent process variables, respectively; $v$ is a vector of process quantities which are a function of $x$; and $p$ is a vector of constant mechanistic parameters. When the defined process is in a transient state, $x$, $y$, and $v$ may vary with time.

This paper considers hybrid modelling scenarios that satisfy the following assumptions:

(a) The form of the function $f$ is known as a consequence of underlying mechanistic knowledge;

(b) The form of the function $h$ is unknown, hence a black-box model requires to be developed to approximate its form; and

(c) $p$ is unknown and requires to be estimated.

Furthermore it is assumed that the structure of $f$ and/or the measurements of $y$ and $x$ are such that, according to Eq. (2), $v$ can be computed analytically or numerically at the sampling points of $x$ and $y$ for a given estimate of $p$:

$$\tilde{v} := f^{-1}(\tilde{x}, \tilde{y}, \hat{p}). \tag{4}$$

It is noted that noisy measurements can pose problems for the computation of $\tilde{v}$. This issue is discussed in the case study in Section 6.

Based on these assumptions, the task of hybrid modelling, as studied in this paper, can be stated as follows: for a given set of measurements, $x$ and $y$, an estimate of $p$ and an approximation of $h$ is obtained such that the resulting model has acceptable capability in terms of predicting the behaviour of the chemical process being modelled.

## 3. A Two-level Solution Approach

To identify the hybrid models described above, a two-level approach, which is an adaptation of the framework for identifying non-parametric (black-box) models with regularization, is developed. Details of the regularized identification of the black-box models are first presented. Extensions to the approach are then proposed to address the estimation of the mechanistic parameters.

### *3.1 Regularized identification of black-box models*

Following the notation defined above, a black-box model is considered:

$$\hat{v} = \hat{h}(x, \theta), \tag{5}$$

where $\hat{h}$ is an estimate of $h$ (defined in Eq. (3)) which represents the form of the black-box model; $\theta$ is the vector of parameters of the black-box model. Under a regularization framework, estimation of the parameters, $\theta$, is achieved through the minimization of a general function that takes the form:

$$I_1 = \sum_{i=1}^{M} c(\tilde{v}_i, \hat{v}_i) + \lambda R. \tag{6}$$

In Eq. (6), the first term on the right-hand side defines the fitness of the model to the training data set, where $M$ is the size of the data set, $\tilde{v}_i$ is measured or derived from the measurements; and $\hat{v}_i$ is the corresponding estimate, $i = 1,\ldots,M$. A common form for this term is the mean or sum of the squared errors. The second term introduces the regularization (or penalty) function, where $\lambda$ is a (weighting) regularization parameter, and $R$ denotes a function of the estimator $\hat{h}$ (or of its parameters $\theta$).

### *3.2 Optimal tuning of the regularization parameter*

Clearly evaluation of the identification criterion as defined in Eq. (6) requires that the regularization parameter $\lambda$ is determined a priori. To select an appropriate value for $\lambda$, a specific criterion or risk function requires to be defined, this is usually an estimate of the

expected generalization error of the model. A large number of the existing criteria cited in the literature belong to the family of cross validation (Craven & Wahba, 1979) and its approximations. A generic form of the criterion for selecting $\lambda$ is denoted as:

$$I_2 = g(\theta). \tag{7}$$

A concrete criterion (leave-one-out Cross Validation) will be given later (cf. Eq. 11).

*3.3 Consideration of the unknown mechanistic parameters*

The preceding sections essentially propose a two-level framework for identifying a black-box model with regularization. For the first level, the parameters $\theta$ are estimated by minimizing $I_1$ in Eq. (6) for a given value of $\lambda$; at the second level, the optimal value for $\lambda$ is computed according to a specific form of Eq. (7). The entire identification task is then completed by iterating between these two levels, until an acceptable result (usually in terms of the value of $I_2$) is attained.

Returning to the hybrid modelling problem under consideration, an additional task is the estimation of the unknown mechanistic parameters $p$ in the mechanistic part of the model (cf. Eq. (2)). To retain the identification framework described above, it is proposed to treat the unknown mechanistic parameters, $p$, as "tuning parameters" in a similar manner to the treatment of $\lambda$. That is, the values are determined at the second level together with $\lambda$ according to Eq. (7), i.e. through the minimization of the expected generalization error. This approach is comparable to that adopted by Wahba and co-workers (Gong et al., 1998) for solving a variational weather prediction problem, where the weighting, smoothing, and mechanistic parameters were simultaneously estimated. However, the problem addressed in their work was that of state estimation (based on a parametric model) as opposed to semi-parametric model identification which is the focus of the research reported in this paper.

It is worth emphasizing that the criterion for tuning the regularization and mechanistic parameters in the setting up of the hybrid model is computed for the *entire* hybrid model, and not just for the black-box component. That is, when generalization performance is considered, it considers the entire hybrid model with respect to the prediction of $y$ in Eq. (2) as opposed to simply focusing on $v$ in Eq. (3).

A schematic of the computational procedure of the two-level approach for solving the problem can now be described (see Figure 1):

a. *Initialization*. Initial guesses for $\lambda$ and $P$ (i.e. $\lambda_0$ and $P_0$) are defined for Levels 1 and 2.

b. *Level 1*: Derive $\hat{v}$ from the measurements of $x$ and $y$ (i.e. $\tilde{x}, \tilde{y}$) at each sampling point according to Eq. (4). Construct a training set from ($\tilde{x}, \hat{v}$). Estimate $\theta$ in the black-box model (Eq. (5)) according to the identification criterion defined in Eq. (6), and pass the resulting estimate of $\theta$ to Level 2.

c. *Level 2*: Evaluate the criterion $I_2$ for selecting $\lambda$ and $P$ according to Eq. (7), adjust $\lambda$ and $P$ by using an optimization algorithm, and pass the new values of $\lambda$ and $P$ back to Level 1. This process continues until the stopping condition of the optimisation algorithm is met. The condition is usually in terms of the maximum number of iterations, the lower limit of the objective function value, etc. A global optimization algorithm would be required if multiple local optima exist.

The convergence property of the approach is essentially embodied within the two-level approach as a consequence of the solvability of the master problem, i.e. the optimization problem at Level 2. In principle, a solution to this problem is obtainable, if the optimization objective (i.e. the measure of model prediction performance) is sensitive to the optimization variables (i.e. the mechanistic and regularization parameters). In comparison with the approach by Weese and co-workers (Weese, 1993; Roths et al., 2001) which co-estimates the parameters of the black-box model and the mechanistic parameters, the above approach identifies the black-

box model alone at Level 1 and hence allows to retain the standard algorithm for this purpose. On the other hand, co-tuning of the regularization and the mechanistic parameters at Level 2 may increase the possibility of encountering multiple local optima as opposed to the tuning of the regularization parameter alone, although the latter alternative has to deal with multiple local optima as well in some circumstances (e.g. Roths et al., 2001).

## 4. Realization of the Two-level Approach

When the identification approach presented in the previous section is realized, a number of technical alternatives exist for the individual steps. In this section, the details of the techniques used in this study for testing the proposed approach are briefly described.

### *4.1 The black-box model and its identification*

As one of a number of non-parametric approaches cited in the literature, the Support Vector Machine (SVM) regression model is considered within this paper due to its simplicity in training (Vapnik, 1999). The SVM has been applied by Yan et al. (2004) and Wan et al. (2005), among others, in process engineering. The SVM model can be defined as follows:

$$\hat{h}(x) = \sum_{i=1}^{P} a_i K(x, x_i), \tag{8}$$

where $a_i$ is an element of the parameter vector, $x_i$ is a vector of independent variables at the *ith* sampling point, *K* is a function referred to as the *kernel* and *P* is the number of parameters or the number of samples included (these two are the same). Where the proximity of the model and the measurements is represented by the sum of squared errors, criterion (6) becomes:

$$I_1 = \sum_{i=1}^{M} (\tilde{v}_i - \hat{v}_i)^2 + \lambda \left\| \hat{h} \right\|^2. \tag{9}$$

By incorporating function (8) into such a criterion and then deriving the optimality conditions, the optimal estimate of the model parameters $a_i$, $i$=1,…,$P$ can be obtained by solving a linear equation system. Details of SVM models and their identification can be found in the literature

(e.g. Vapnik, 1999). Of a number of possible alternatives, *K* is defined as the radial basis function:

$$K(x_i, x_j) = \exp\left(\frac{-\left\|x_i - x_j\right\|^2}{2\sigma^2}\right), \tag{10}$$

where σ is a real-valued parameter of the kernel. To finally determine model (8), the value of σ requires to be defined. It is reasonable to assume that through the determination of σ, the expected generalization error of the hybrid model of which the SVM model is the black-box component should be reduced. Therefore, the approach adopted is to tune σ at the second level of the identification framework presented earlier, i.e. determine σ together with the regularization parameter λ and the mechanistic parameters *P*.

*4.2 Tuning of level-2 parameters: Criterion and optimization algorithm*

To demonstrate the approach, leave-one-out cross validation (CV) was selected as the criterion for tuning the parameters at Level 2 due to its simplicity of implementation:

$$I_2 = \sum_{i=1}^{M} (\tilde{y}_i - \hat{y}_i(\hat{a}_{-i}))^2, \tag{11}$$

$\tilde{y}_i$ is the *ith* measurement; $\hat{a}_{-i}$ is the vector of parameters in function (8) and is estimated using a training data set which excludes the *ith* measurement; and $\hat{y}_i(\hat{a}_{-i})$ is the prediction of the *ith* measurement based on $\hat{a}_{-i}$. This implies that each execution of step (b) in Figure 1 materializes in the identification of the SVM model with *M*-1 samples, and this step is executed *M* times to accomplish one execution of step c(i).

The parameter tuning problem at Level 2 is essentially one of multi-variable, non-linear optimization. Preliminary studies show that there exist multiple minima of the CV tuning criterion. To deal with local minima, the Generic Algorithm (GA) was selected as the optimizer. The SVM-CV realization was undertaken in MATLAB. SVM and CV are self-implemented.

For the optimizer at Level 2, a GA package that implements the Differential Evolution algorithm (Price, Storn, 1996) was employed.

## 5. Case Study: Modelling of the Toluene Nitration Process

Toluene nitration is a heterogeneous liquid-liquid batch reaction process with a number of possible operational regimes. It has been studied previously by several authors and its mathematical model has been presented in the literature (e.g. Zaldivar et al, 1995). In this study, this process is selected for demonstrating, through simulation studies, the proposed hybrid modeling approach which involves (i) the identification of a SVM model for the chemical reaction kinetics, and (ii) the determination of the mass transfer parameters. In the subsequent sections, the model sections relevant to the current study and how the training/test data sets are prepared through simulation are described.

### *5.1 The hybrid model structure*

The model corresponding to the slow reaction regime of the nitration reactor is considered:

$$\frac{dn_{toluene}}{dt} = -r\, V_{reaction}, \tag{12.1}$$

$$r = \frac{m\ C_{toluene,organic}}{\frac{1}{k_L a} + \frac{1}{k(1-\varphi)C_{HNO_3}}}, \tag{12.2}$$

$$a = \frac{6\varphi}{dA(1+B\varphi)We^{Q}}. \tag{12.3}$$

In Eq. (12.1), $n_{toluene}$ is the mole number of toluene in the reactor, $V_{reaction}$ is the volume of the entire reaction medium; and $r$ is the overall conversion rate. In Eq. (12.2), $C_{HNO3}$ is the molar concentration of $HNO_3$; $C_{toluene,organic}$ is the molar concentration of toluene in the organic phase; $a$ is the specific surface area for inter-phase mass transfer; $\varphi$ is the volume ratio of the organic phase; $k$ is the apparent reaction rate constant; $k_L$ is the mass transfer coefficient of toluene; and $m$ is the distribution coefficient of toluene. In Eq. (12.3), $d$ is the diameter of the stirrer, $We$ is

the Weber number of the continuous phase (as a function of the stirrer speed), and $A$, $B$, and $Q$ are regression parameters for calculating the Sauter mean diameter of the dispersed liquid droplets (known as $d_{32}$). Considering that the amount of toluene in the organic phase is dominant (in comparison with that in the aqueous phase) and that the change of overall reaction volume ($V_{reaction}$) is insignificant during an experiment, simple manipulation of Eq. (12.1) leads to:

$$\frac{d(C_{toluene,organic}\varphi)}{dt} = -r. \tag{12.1a}$$

Furthermore, the apparent reaction rate constant $k$ (in Eq. (12.2)) is a complex function associated with the conditions of the reaction including its temperature and the composition of the mixed acid in the aqueous phase. In this study, it is assumed that this function is unknown and therefore requires to be approximated through a black-box model. Since the data used for training are all based on isothermal reactions, the effect of temperature on $k$ is eliminated, hence:

$$k = h(C_{HNO_3}, C_{H_2O}, C_{H_2SO_4}). \tag{12.4}$$

Finally, with regard to the mechanistic parameters, it is assumed that $k_L$, *m, A, B and Q* may be unknown and thus require to be estimated. All quantities other than these parameters and $k$ will become available either directly by measurements or by derivation from measurements; further details are given in Section 5.2. As such, the structure of the hybrid model to be identified becomes clear: Eqs. (12.2) and (12.4) correspond to Eqs. (2) and (3), respectively.

*5.2 Preparation of data for training and testing*

To identify the model defined in Eq. (12.4), a training data set is required in which each sample comprises three independent variables, $C_{H2O}$, $C_{HNO3}$, $C_{H2SO4}$ and the dependent variable, $k$. The three concentrations can be measured, while $k$ has to be derived from the measurements by applying Eq. (12.2). According to Eq. (12.2), $k$ can be computed via a realization of Eq. (4) provided that all other variables/parameters in this equation are known. The situation is: (a) $C_{toluene,organic}$ and $C_{HNO3}$ are measured; (b) $We$ and $\varphi$ can be derived from the measurements and

the mathematical model at the sample points; (c) parameters $k_L$, *m, A, B, and Q* require to be given a value before Eq. (12.4) is identified according to the two-level approach (cf. Figure 1); and finally (d) *r* is derived from the measurements of $C_{toluene,organic}$ according to Eq. (12.1a).

Point (d) requires to be addressed carefully as the noise in the measurement of $C_{toluene,organic}$ can be significantly amplified when computing *r* if no appropriate measure is taken. In this study, the smoothing cubic spline technique (Reinsch, 1967) is applied to smooth the noisy data. This results in a smooth time curve of the quantity ($C_{toluene,organic}\ \varphi$), which is then analytically differentiated to obtain an estimate of *r* (cf. Eq. (12.1a)). Similar treatments have been applied in a number of chemical kinetics modelling studies that follow a differential identification approach (e.g. Yeow et al., 2003; Bardow & Marquardt, 2004).

The data that can be measured or derived from measurements are generated through numerical simulation using the simulator gPROMS (Process Systems Enterprise, 2004), based on a rigorous model of the nitration process reported in Zaldivar et al (1995). Random noise was added to the values of $C_{toluene,organic}$ to mimic the real process measurements. Three measurement error levels were applied to investigate model identification performance, they corresponded to a relative error of 0 (i.e. error-free), 0.5, and 5 percent (in terms of the standard deviation of a Gaussian distribution). Applying the aforementioned treatments, four simulation runs were performed with different stirrer speeds (*N*), each of which was of 400 minutes duration and these were sampled every 2 minutes. The value of *N* was set to 10 $s^{-1}$, 12 $s^{-1}$, 15 $s^{-1}$, and 13.5 $s^{-1}$, respectively, while all other operating conditions were kept the same. The first three simulations collectively generated data for identifying the model (random noise at three levels was added to the data to obtain three training data sets), whilst the last one was for the testing of the model (hence no noise was added).

Figure 2 shows, as an example, the result of processing the noisy concentration data using cubic smoothing splines to obtain the data of the overall conversion rate when $N = 12\ s^{-1}$. Due to the poor performance at the boundaries, which is a well known problem (Haerdle, 1990), 40 points at each end of every curve of the overall conversion rate $r$ were discarded. Furthermore, the relatively high sampling rate (one sample per 2 minutes) has been adopted primarily to attain accuracy with respect to deriving $r$ through the use of cubic smoothing splines. The computational load of the identification process increases when the size of the training data set becomes larger. Preliminary studies showed that a training data set with a size of 60 samples would be sufficient. Therefore after the above derivation was accomplished, only one sixth of the data (after the boundary points were truncated) for each given conversion rate, $N$, was used in a training data set. Consequently, the size of one entire training data set was (400/2 – 40 *2)/6*3 = 60. This allowed to prepare six distinct training data sets. Each data set was subsequently used for performing one set of numerical experiments of model identification, leading to six sets of repetitive numerical experiments in total.

## 6. Results and Discussion

The identification of the hybrid model was conducted with training data at three different levels of measurement errors as described above. All predictions were made for the batch where $N$ = 13.5 (i.e. the one that generated the test data set). Six repetitive sets of numerical experiments were performed; the mean values and the standard deviations of the estimates of the mechanistic parameters were calculated. Each set of numerical experiments included two different groups of identification studies. In the first group each identification run assumed that only one mechanistic parameter is unknown. The results of the parameter identification are shown in Table 1. Figure 3 shows the performance of the predictions of the resulting models when applied to the test data set. Only the performance of the models in which $m$ was estimated are shown

here; other models exhibited similar performance. It is noted that in all the figures in this section, a plot of "measurements" is always based on the simulation data without added noise.

In the second group, the simultaneous estimate of the multiple mechanistic parameters was studied. For this group, the issue of identifiability needs to be considered. Combining Eqs. (12.2) and (12.3) gives:

$$r = \frac{m\ C_{toluene,organic}}{\frac{dA(1+B\varphi)We^{Q}}{6\varphi k_L} + \frac{1}{k(1-\varphi)C_{HNO_3}}}, \tag{13}$$

Eq. (13) can be further rewritten as:

$$r = \frac{C_{toluene,organic}}{\frac{dA'(1+B\varphi)We^{Q}}{6\varphi} + \frac{1}{Sk(1-\varphi)C_{HNO_3}}}, \tag{14.1}$$

$$A' = A/(mk_L), \tag{14.2}$$

$$S = mk_L. \tag{14.3}$$

It is evident from Eq. (14.1) that $S$, as an unknown constant parameter, cannot be uniquely identified when the black-box model of $k$ also requires to be identified, because there is no means to separate the influence of $S$ and that of $k$ on the model. Thus, whilst five mechanistic parameters are present when the model is written in the form of Eq. (13), rewriting this equation in the form of Eq. (14) reveals that only three parameters ($A'$, $B$, and $Q$) are identifiable. Denoting

$$k' = Sk, \tag{14.4}$$

the above discussion implies that it is more appropriate, from the identifiability perspective, to formulate the problem of identifying the hybrid model under consideration as one that estimates simultaneously parameters ($A'$, $B$, and $Q$) and at the same time identifies a black-box model of $k'$.

Furthermore, in previous modelling studies, (Zaldivar et al, 1995)‡, it has been recognized that the estimate of $B$ in Eq. (13) has an insignificant impact on the prediction error when $A$ and $B$ are estimated simultaneously. This is hypothesised to be due to the insignificant change in $\varphi$ (usually within 10%) throughout a reaction batch. The same situation exists in Eq. (14). To investigate the implication of the case of identifiability, two studies were performed in this second group, one where the three parameters ($A'$, $B$, and $Q$) were co-estimated, whilst the other assigned the true value to $B$ and co-estimated the other two parameters. The results of parameter identification for these two studies (both belonging to the second group) are shown in Tables 2 and 3. Figure 4 shows the performance of the prediction of the resulting models when applied to the test data set.

The following observations can be made on the basis of the above model identification results:

(1) In the ideal cases where no measurement errors are present and when the identifiability is fair, , the identification of the semi-parametric hybrid model using the two-level approach yields perfect estimates of the unknown mechanistic parameters (cf. the rows corresponding to "*Using error-free training data"* in Tables 1 and 3), which can be considered as comparable to the cases of parametric model identification.

(2) The estimates of unknown mechanistic parameters are still fairly accurate when small measurement errors are present, although the accuracy decreases when the measurement errors become larger. However, even in such cases, the quality of model prediction is still not significantly impaired.

It is worth noting that, when the measurement errors are larger in magnitude, the estimates of the mechanistic parameters become less reliable, but the hybrid modelling paradigm can still be advantageous over that of a purely black-box modelling approach, particularly in terms of

‡ Due to this reason, a value of "2.0" was assigned to $B$ in Zaldivar et al (1995), while the value of $A$ was estimated from the experimental data. The parameter set generated this way was actually taken for running stimulations in this study.

extrapolation. To illustrate this point in the context of this case study, a purely black-box SVM model for predicting the overall conversion rate $r$ was built. This model maps six input variables, namely $C_{toluene,organic}$, $C_{H2O}$,$C_{HNO3}$, $C_{H2SO4}$, $\varphi$, and $We$ to the output variable $r$. Only the training data set with a measurement error of 5% was used to train the SVM model, but the data set originally used in the hybrid modelling was augmented by doubling the sampling rate, consequently twice the number of data points were used. The regulation parameter $\lambda$ and the kernel function parameter $\sigma$ were determined using the Leave-one-out cross validation method and the, same GA optimization tool as applied for tuning the Level-2 parameters in the hybrid modelling.

Figure 5 shows how this pure black-box SVM model performs when applied to the test data set (i.e. the one generated with $N = 13.5$). It can be observed that it is comparable to that of the hybrid model (cf. Figure 4, dotted curves). However, such a prediction is basically an interpolation of the training data sets which, as mentioned earlier, comprise data generated from $N = 10$, $N = 12$, and $N = 15$ whilst all other conditions are the same.

The resulting SVM model was then further tested with another set of data, which was again generated with $N = 13.5$ but with a 5% increase in the initial amount of toluene in the reactor compared with the conditions applied in all previous trainings/tests. The major influence of this increase is on the range of the concentration of toluene in the organic phase ($C_{toluene,organic}$) and that of the volume ratio of the organic phase ($\varphi$), both being an input to the SVM model. This is essentially an extrapolative prediction of the models, because the ranges of independent variables for the prediction are beyond those of the data applied for training the models. Figure 6 shows a comparison of the prediction performance of the SVM model and that of the hybrid model identified earlier which has three estimated parameters, $A'$, $B$, and $Q$.

The poor performance of the SVM model can be explained as a consequence of a black-box model generally not being applicable for extrapolation. In contrast, the comparative results

shown in Figure 6 demonstrates the advantages of a hybrid model in this situation. In general, such a desirable capability, i.e. reliable extrapolation using a hybrid model, is most likely to occur when the extrapolated conditions are not part of and/or have a minor influence on the inputs of the black-box component of the hybrid model. This is the case for this current extrapolation test, where the two quantities, $C_{toluene,organic}$ and $\varphi$, whose ranges were notably changed, are inputs to the SVM model but not the SVM part of the hybrid model. This indicates that constructing a semi-parametric hybrid model is still preferable to that of a pure black-box model, even when the estimates of the mechanistic parameters cannot be sufficiently accurate due to the measurement errors in the training data set. Such a semi-parametric hybrid model can be built using the approach developed in this work.

## 7. Conclusions

Hybrid models are being more widely considered and applied for the modelling of chemical processes. In this paper, semi-parametric hybrid models which combine a mechanistic component with unknown parameters and a non-parametric black-box component have been investigated. The construction of such a hybrid model involves simultaneously the estimation of the unknown mechanistic parameters and the identification of the non-parametric function (i.e. the black-box part). A two-level approach is proposed. It decomposes the modelling task into (i) the regularized identification of the black-box component of the entire model at the first level and (ii) the optimal tuning of the regularization parameters and the mechanistic parameters simultaneously at the second level. The overall goal of identification is to attain the best generalization performance of the hybrid model as a totality.

The evaluation of this approach, which utilised the Support Vector Machine (SVM) for the black-box modelling component and cross validation as a measure of generalization performance, was demonstrated on a previously studied toluene nitration modelling problem.

The results showed that the proposed model identification framework has the potential to accurately estimate the mechanistic parameters in a hybrid model when the measurement error in the training data is small. When the measurement error increases in magnitude, the estimates of the mechanistic parameters reduce in accuracy. However, acceptable prediction performance of the resulting model can still be obtained, and the hybrid model continues to outperform the pure black-box model under extrapolation circumstances, particularly when the extrapolated conditions are not part of or have a minor influence on the inputs of the black-box component of the hybrid model.

**Acknowledgments**

The authors acknowledge the support of the EPSRC award GR/R64407/01 "Vertical Integration of Product Development and Manufacturing".

**References**

Bardow, A., Marquardt, W. (2004). Incremental and simultaneous identification of reaction kinetics: methods and comparison. *Chem. Engng. Sci*. 59, 2673–2684.

Chen, L., Bernard, O., Bastin, G., Angelov, P. (2000). Hybrid modelling of biotechnological processes using neural networks. *Control Engineering Practice*, 8, 821–827.

Craven, P., Wahba, G. (1979). Smoothing noisy data with spline functions. *Numer. Math*. 31, 377–403.

Duarte, B., Saraivay, P. M., Pantelides C. C. (2004). Combined Mechanistic and Empirical Modelling. *Int. J Chem. React. Eng.*, Vol. 2, Article A3.

Eubank, R.L. (1988). *Spline Smoothing and Non-parametric regression.* Marcel Dekker, INC., New York.

Haerdle, W., Mueller, M., Sperlich, S., Werwatz, A. (2004). *Non-parametric and semiparametric models*. Springer, Berlin.

Hastie, T.J., Tibshirani, R.J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Houck, C., Joines, J., Kay, M. (1995). *A Genetic Algorithm for Function Optimization: A MATLAB Implementation*. NCSU-IE TR 95-09. On-line available at http://www.ie.ncsu.edu/mirage/GAToolBox/gaot/papers/gaotv5.ps, accessed 04.01.2010.

Geman, S., Bienenstock, E., Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4, 1-58.

Gong, J., Wahba, G., Johnson, D. R., Tribbia, J. (1998). Adaptive Tuning of Numerical Weather Prediction Models: Simultaneous Estimation of Weighting, Smoothing and Mechanistic parameters. *Monthly Weather Review*, 126, 210.

Green, P.J., Silverman, B.W. (1994). *Non-parametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman and Hall, London.

Kahrs, O., Marquardt, W. (2008). Incremental identification of hybrid process models, *Computers & Chemical Engineering,* 32, 694-705.

Lima, P., V., Saraiva, P. M., and GEPSI-PSE Group (2007). A semi-mechanistic model building framework based on selective and localized model extensions. *Computers & Chemical Engineering*, 31, 361-373.

Montague, G., Morris, J. (1994). Neural network contributions in biotechnology. *Trends in Biotechnology*, 12, 312–324.

Oliveira, R. (2004). Combining first principles modelling and artificial neural networks: a general framework. *Comp. Chem. Engng*. 28, 755–766.

Price, K., Storn, R. (1996). MATLAB code of the Differential Evolution algorithm. On-line available at http://www.icsi.berkeley.edu/~storn/code.html, accessed 05.12.2005.

Process Systems Enterprise (2004). *gPROMS Advanced User Guide*, Process Systems Enterprise Ltd., 23.02.2004.

Psichogios, D. D., Ungar, L. H. (1992). A hybrid neural network—First principles approach to process modelling. *AIChE Journal*, 38(10), 1499–1511.

Reinsch, C. (1967). Smoothing by spline functions. *Numer. Math*. 10, 177-183.

Roths, T., Marth, M., Weese, J., Honerkamp, J. (2001). A generalized regularization method for nonlinear ill-posed problems enhanced for nonlinear regularization terms. *Computer Physics Communications* 139, 279-296.

Ruppert, D., Wand, M.P., Carroll, R.J. (2003). *Semiparametric regression*. Cambridge University Press, Cambridge.

Rusinowski,H., Stanek, W. (2009). Hybrid model of steam boiler, *Energy*, in press.

Saraceno, A., Curcio, S., Calabro,V., Iorio, G. (2009). A hybrid neural approach to model batch fermentation of 'ricotta cheese whey' to ethanol, *Computers & Chemical Engineering*, in press.

Teixeira AP, Carinhas N, Dias JML, Cruz P, Alves PM, Carrondo MJT, Oliveira R (2007). Hybrid semi-parametric mathematical systems: Bridging the gap between systems biology. *Journal of Biotechnology* 132 ,418-425.

Thompson, M. L., Kramer, M. A. (1994). Modelling chemical processes using prior knowledge and neural networks. *AIChE Journal*, 40(8), 1328–1340.

Tian,Y., Zhang, J., Morris, J. (2001). Modelling and Optimal Control of a Batch Polymerization Reactor Using a Hybrid Stacked Recurrent Neural Network Model. *Ind. Eng. Chem. Res*., 40, 4525-4535.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer-Verlag, New York.

Wan, X., Pekny, J.F., Reklaitis, G.V. (2005). Simulation-based optimization with surrogate models—Application to supply chain management. *Comp. Chem. Engng*. 29, 1317-1328.

Wang,X., Chen,J., Liu,C., Pan,F. (2009). Hybrid modeling of penicillin fermentation process based on least square support vector machine, *Chemical Engineering Research and Design*, in press.

Weese, J. (1993). A regularization method for nonlinear ill-posed problems. *Computer Physics Communications* 77, 429-440.

Yan, W., Shao, H., Wang, X. (2004). Soft sensing modeling based on support vector machine and Bayesian model selection. *Comp. Chem. Engng*. 28,1489-1498.

Yeow Y. L., Wickramasinghe, S. R., Hanb, B., Leong, Y.-K. (2003). A new method of processing the time-concentration data of reaction kinetics. *Chem. Engng. Sci.* 58, 3601–3610.

Zaldivar, J.M., Alós, M., Molga, E., Hernández, H., Westertep, K.R. (1995). Aromatic nitration by mixed acid: slow liquid–liquid reaction regime. *Chem. Eng. Process.* 34, 543–559.

Figure 1.  A two-level approach for the identification of hybrid models.

Figure 2. Derivation of overall conversion rates using smoothing cubic splines ($N=12s^{-1}$).

Figure 3. Prediction performance of models identified in the first group (with *m* as the estimated mechanistic parameter).

Figure 4. Prediction performance of models identified in the second group.

Upper plot: co-estimation of three mechanistic parameters.

Lower plot: co-estimation of two mechanistic parameters.

Figure 5. Interpolation performance of the black-box SVM model.

Figure 6. Extrapolation performance of the hybrid model and the black-box model.

| Identification case | Mechanistic parameter | | | | |
|---|---|---|---|---|---|
| | *m*1e4* | $k_L$**1e5* | *A* | *B* | *Q* |
| *True value* | 4.000 | 1.660 | 0.3512 | 2.000 | -0.6000 |
| *Using error-free Training data* | 4.000 ±1.971e-5 | 1.660 ±7.313e-6 | 0.3512 ±1.949e-6 | 2.000 ±6.363e-5 | -0.6000 ±7.488e-7 |
| *Using training data with 0.5% measurement error* | 3.909 ±5.260e-4 | 1.622 ±2.188e-4 | 0.3614 ±1.923e-4 | 2.240 ±1.600e-3 | -0.5960 ±1.938e-5 |
| *Using training data with 5% measurement error* | 3.595 ±1.887e-2 | 1.492 ±7.009e-3 | 0.4004 ±2.600e-3 | 3.149 ±5.080e-3 | -0.5815 ±9.097e-3 |

Table 1. Parameter identification results of the studies in the first group.

| Identification case | Mechanistic parameters | | |
|---|---|---|---|
| | *A'*1e7* | *B* | *Q* |
| *True value* | 5.289 | 2.000 | -0.6000 |
| *Using error-free training data* | 5.289 ±0.3309 | 2.630 ±1.481 | -0.6088 ±0.0184 |
| *Using training data with 0.5% measurement error* | 9.364 ±1.995 | 0.4643 ±0.4612 | -0.6628 ±0.0360 |
| *Using training data with 5% measurement error* | 5.808 ±1.817 | 0.3002 ±0.0845 | -0.5599 ±0.0460 |

Table 2. Parameter identification results of the studies in the second group: co-estimation of three mechanistic parameters.

| Identification case | Mechanistic parameters | |
|---|---|---|
| | *A'*1e7* | *Q* |
| *True value* | 5.289 | -0.6000 |
| *Using error-free training data* | 5.287<br>±4.255e-3 | -0.5999<br>±1.147e-4 |
| *Using training data with 0.5% measurement error* | 6.549<br>±5.791e-1 | -0.6258<br>±1.280e-2 |
| *Using training data with 5% measurement error* | 5.224<br>±2.093 | -0.5728<br>±4.960e-2 |

Table 3. Parameter identification results of the studies in the second group: co-estimation of two mechanistic parameters.