Model Selection: an overview of practices in chemical engineering

Peter J.T. Verheijen

Dept. of Chemical Engineering, Technological University Delft, Julianalaan 136, 2628 BL Delft, The Netherlands, e-mail: P.J.T.Verheijen@tnw.TUDelft.nl

Abstract

The problem of choosing out of a large set of models is considered here. Five criteria are mentioned: the level of rigor that a model describes phenomena; the accuracy of the model with respect to data; the adequacy of the model for the purpose of simulation, optimization or design; the flexibility of a model; and the computational complexity. The focus is on the purely statistical criteria. A systematic procedure is presented to eliminate sequentially hierarchical models and equivalent models, based on the F-test and Bartlett's χ^2-test. The use of Bayesian statistics as a tool is reviewed. And recent developments in optimization-based approaches are listed. An overview of available measures for the distance between models is given. It is shown that one should not accept a single model, as is engineering practice, but rather allow for a set of models when the selection does not give sufficient reason to discriminate between models. Examples from the area of reaction kinetics and adsorption are given.

1 Introduction

After engineering insight in a process, the most important step is to be able to make a quantitative model of that process in order to test the understanding gained, or to make use of the model in engineering decisions for design or operation.

In the area of kinetics, the typical problem is the level of detail required in the reaction network. Normally, different networks are proposed, that make sense chemically, and a choice must be made on the basis of the available data. The thermodynamic model of a multi-phase, multi-component system is primarily determined by the attributes, e.g. polar or non-polar, of the molecules involved, but also different models are often considered acceptable by the experts. Individual process units can be governed by very simple short-cut models or very complex detailed rigorous models. An example is the heat

Article published in Computer-Aided Chemical Engineering 16 (2003) 85–104

exchanger described either by one simple area and one heat coefficient, or by a detailed CFD-description accounting for the flow details. Another example is the multi-compartment model of a tank based process unit, such as a crystallizer. The number of compartments is a "design" variable of the model.

A second dimension is the use of the model. The science view is that a model should be sufficiently complex, to be sure that all phenomena are described. Scientific work is namely aimed at understanding fundamental phenomena. The engineering view is that a model should be sufficiently simple to describe only the really necessary phenomena. This sets different requirements on the level of complexity. Especially in the latter case, the model user can put limits on the operation ranges — temperature and pressure — in order to enhance insight for design or for computability.

In practice it is often difficult to generate many models and the theory often limits the number of possibilities. So, this reduces a potentially complex to one of a simple choice between two of three competing models, and the engineer of scientist only wants to make a choice.

A generic approach to model selection is presented here to encompass the whole selection process. The problem statement is: *Assume, that data, a set of models, and a purpose for the model is given. Here we present methodologies for the selection of a sub-set of models or a single model that fulfils the model purpose and that is consistent with the data.*

At the root of this process, lies the need to fit the model to the data. So, it is also assumed that some parameters, β, are to be estimated from the data, Y. The model is the functional relation, $f(\mathbf{x};\beta)$, expressing one or more observables as a function of set point, $\mathbf{x}$, and a set parameters, β, to be estimated from an experiment. Next to the model and the data also the "error model", i.e. the width of the error distribution, σ_i, should be known at each measurement, $\mathbf{x}_i$. The estimate of the parameters comes from the usual minimization (e.g. Bates and Watts [1]) of

$$SS_{\text{res}}(\beta) = \sum_i \left(\frac{Y_i - f(\mathbf{x}_i;\beta)}{\sigma_i} \right)^2 . \tag{1}$$

with respect to the parameters, β. The parameter estimate, $\hat{\beta}$, its covariance matrix, the value of the minimum sum squared, $SS_{\text{res}}(\hat{\beta})$, and the calculated responses, $f(\mathbf{x}_i;\hat{\beta})$, are the direct quantitative results.

Essentially, we will follow three approaches. Firstly, an inference approach reduces the whole process to a sequence of decisions. The end result is a clear accept or reject for the candidate models. Secondly, the selection problem can be restated as an optimization problem. The selection process reduces to the

obtainment of a single candidate model that is found to reach an optimum for some chosen objective. Thirdly, a more subtle evaluation of each alternative is given, by calculating the probability for each of the candidate models and choosing the model with the largest probability given the data.

2 Criteria

In the whole process from model to a quantitative support for analysis and design the following can be distinguished:

Physico-chemical criteria are essential. The phenomenon should be described by the right choice of parameters, variables and equations.

Flexibility of a model is interesting in respect to its re-use in various situations. The re-use not only makes the model more useful, but also implicitly tests the underlying ideas.

Computational criteria describe the process of computing a model within a reasonable time and within limited memory resources.

Statistical criteria evaluate the distance between the model and data. This is governed by the uncertainty of the measurements. In more complex situations also the actual statistical distribution of the data have to be taken into account.

Engineering criteria refer to the usefulness of the model. The model should describe the process well enough within the range specified.

These criteria are contradictory. In engineering practice, the expedient and direct usefulness is considered more important than any of the other criteria. In a later stage, the modeler then often finds that the whole process has to restart from the very beginning. The investment in models, whose rigor, flexibility and computability have been optimized, is much larger, but it has a long-term benefit.

Once the degree of rigor, the range of complexity is set, and the models can be computed, only the statistical criteria guide with the remaining questions of adequacy and selection.

Linhart and Zucchini [2] give a summary of statistically relevant measures to express the distance between model and data. Uncomplicated measures, which are especially found in the systems identification literature [3], are based on

the sum of squares in its various forms,

$$\begin{array}{lll} \text{Error variance} & s^2_{\text{res}} & = SS_{\text{res}}/(n-p) \\ \text{Akaike's Final Prediction Error} & FPE & = s^2_{\text{res}} \frac{n+p+1}{n-p-1} \\ \text{Akaike's Information Criterion} & AIC & = n \ln(s^2_{\text{res}}) + 2p \\ \text{Shortest Data Descriptor} & SDD & = \ln(s^2_{\text{res}}) + (p+1)\ln(n). \end{array} \tag{2}$$

The simplest and most powerful remains the error variance. The final prediction error and the shortest data descriptor are both designed to favor models with less parameters. Their usefulness lies in the fact that they emphasize the robustness of the predictions done with the models. Akaike's information criterion more closely follows the scientific paradigm to retain information. This gives models with a higher number of parameters an advantage.

A second set of criteria is based on a general norm: $L_1, L_2, \ldots L_\infty$, where of course the L_2 norm is related to sum-of-squares criteria above. These norms are useful in parameter estimation as they represent difference distance measures, which can be adjusted to suit circumstances. In model selection they are less useful as an interpretation of the criteria is lacking. Also, the higher-order norms emphasize the larger residuals, and are therefore comparing different models on just a few locations in the response space, rather than the models in a specified domain.

A third set of criteria is based on statistical measures. The likelihood or the log-likelihood function is directly related to the maximum likelihood estimator, and it is also related to the Akaike's information criterion [3]. Lastly, the Bayesian criterion of posterior model is in fact the model probability, which is the subject of a separate section below.

The pragmatic approach is to take one of the measures in eq. (2).

3 Inference approach

Statistical inference starts from a hypothesis and relies on a framework to come to a clear decision. The hypothesis is rejected, or is not rejected. In order to apply this to the model selection issue, the whole process is to be divided in stages. Here, we will define three such stages. In the first stage, each model is compared to the data, and tested on adequacy. The second stage identifies the common situation that there exists sets of models related to each other but only differing in complexity. Here, tests of model reduction to choose an appropriate level of complexity can be performed. The third stage puts the

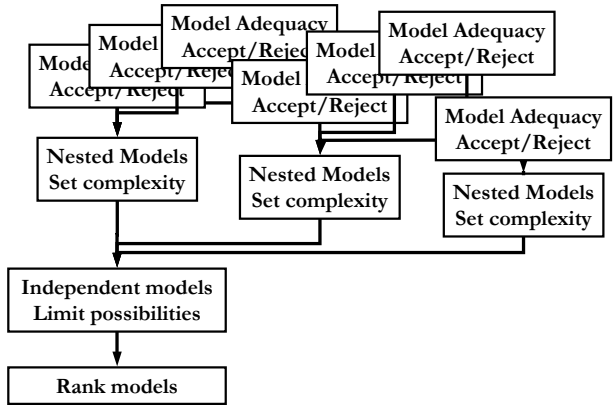

Fig. 1. Decision tree for selecting models by inference. The four levels are treated in sections 3.1, 3.2, 3.3 and 3.4, starting from the top

remaining models together and seeks a test that further reduces this set. The top three levels of Figure 1 illustrate the three stages.

3.1 Model adequacy

We are considering here the process of rejecting or accepting a single given model with the given data. Physico-chemical insight should be the first criterion to be used. Even a simple polynomial approximation assumes that the response variable fulfils continuity properties. However, this leads to the general question of model validation which also on the specific discipline, e.g. Murray-Smith [4]. Simple actions are:

- Visual inspection of particular responses before and after fitting the data. The opinion of experts, or the response of knowledge management systems.
- Degeneracy tests, i.e. defining limiting cases such as extremely high and low temperatures and simulate the model in those regions.
- Comparison tests, where model simulations are contrasted to e.g. short-cut models.

If we further assume, that the model at least fulfils its engineering purpose, there remain statistical measures to consider.

Known measurement errors allow the well known χ^2-test.
Repeated measurements are the basis for the lack-of-fit test.
Residuals can be tested whether they reject the assumed distribution, often the normal distribution.
Measurements in time-series can be tested on independence between residuals such as applied in system identification [3].

In the last decade, it is becoming accepted practice to reconcile data [5] with the constraint that mass balances should always be fulfilled. This leads to better estimates at the measurement positions, but this process allows also for intermediate lack-of-fit testing, as the mass balance hypothesis can be tested as well.

Ideally speaking, a model should be rejected when it fails one of the statistical tests. However, a model is often accepted in practice, especially when the alternatives are not much better. In that case, it should be realized that the statistic defined in eq. (1) has no statistical meaning, and is nothing more than a distance measure of equal footing with any other distance measure, especially the L_2-norm.

3.2 Nested models

In this case, between the models exist a hierarchy such that one model is a special case of an extended model. The obvious example is the class of polynomials, where the appropriate degree should be found. The Padé approximations are the natural extension of the polynomial class, but they are often overlooked. Within each class a smaller model can be seen as a simplification of a more complex model. It is a single power term that can be added or deleted [1] The virial expansion in thermodynamics has a similar property, and in kinetics it is the gradual increase of the reaction scheme that makes it fortuitous to pose these as nested models as well.

Here we consider one class of nested models, and the method is to come to one single candidate model by the method of continuous elimination by inference. In the engineering approach — aiming for simple models — the fits of two nested models can be compared. We start from the simplest model and add terms until the statistic

$$T = \frac{(SS_{\mathrm{res}}(\hat{\beta}(p+p_1)) - SS_{\mathrm{res}}(\hat{\beta}(p)))/p_1}{SS_{\mathrm{res}}(\hat{\beta}(p+p1))/(n-p-p_1)} \sim F_\alpha(p_1, n-p-p_1), \qquad (3)$$

is inferred not to reject the last outcome with the given significance level α. Here p is the number of data of the lower order model, p_1 are the added number of parameters, and n is the number of data.

In the science approach — aiming for the complex models — we start from the most complex model that is commensurate with the given number of data.

[1] Note that the class of polynomials is a special case of the Padé approximations.

The statistic

$$T = \frac{(SS_{\text{res}}(\hat{\beta}(p+p_1)) - SS_{\text{res}}(\hat{\beta}(p)))/p_1}{SS_{\text{res}}(\hat{\beta}(p))/(n-p)} \sim F_\alpha(p_1, n-p), \tag{4}$$

is evaluated.

The result of each this process is a single model for each set of nested models.

3.3 Independent models

Each of the classes of nested models, or each stand-alone model that cannot be related to any other model is considered as independent or non-nested models. The result of the fit process, SS_{res}, for each of the models should have a known distribution, the χ^2-distribution after scaling with the yet unknown error variance, σ^2.

The assumption of complete independence of this set of models is a rather abstract and not a pragmatic proposition. However, with this assumption Bartlett's χ^2-test of homogeneity of variances can be applied. Namely,

$$T = \frac{\sum_{m=1}^{M}(n-p_m)\ln(s^2_{\text{tot}}/s^2_m)}{1+\frac{1}{3(M-1)}\left[\sum_{m=1}^{M}\frac{1}{n-p_m} - \frac{1}{\sum_{m=1}^{M} n-p_m}\right]} \sim \chi^2(M-1), \tag{5}$$

where M is the number of models, p_m is the number of parameters for each model m, s^2_m is the error variance of model m, and s^2_{tot} the error variance based on all error variances together. Eq. (5) can be evaluated and if it leads to rejection the model associated with the largest error variance is deleted from the set. This process is repeated until a set is obtained which does not fail the inference test. It can be reasoned that this elimination process is not exhaustive but, more importantly, it will not eliminate the most appropriate model.

The result is small set of models that cannot further be distinguished.

3.4 Decision process

There are then three clearly defined stages in the inference approach. The model adequacy stage allows considering physico-chemical criteria, and leads to a set of models and classes of models. The model reduction phase with

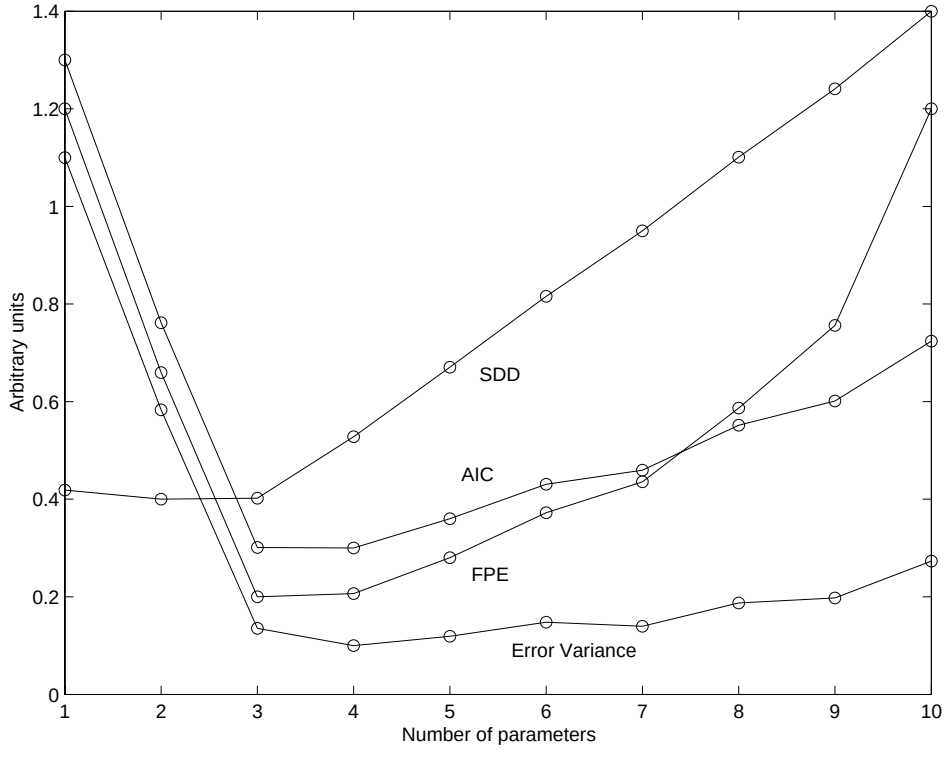

Fig. 2. A hypothetical case of 11 data points fitted with models containing from 1 to 10 parameters. The simulation data were based on a model containing 4 parameters.

nested models allows the introduction of engineering criteria. Choosing one model within each class further reduces the model set. At the end the remaining independent models can be further restricted by mutual comparison.

If still some candidate models remain, there is no further possibility to come to a single model. It remains to rank the models with respect to some measure. The measures mentioned in eq. (2) are the preferred candidates, as they are easily calculated. Figure 2 illustrates this schematically. It is clear that the shortest data descriptor favors a model with less parameters, while the error variance and AIC tend to the more complex models.

The structure of the three decision stages with the final ranking is shown in Figure 1.

4 Optimization approach

A more forthright approach is to define a criterion such as given in section 2, and simply choose the model that has an optimum for the given criterion. The ranking given in the previous section is in fact an optimization applied

after the set of models is reduced by inference. The extension is to evaluate the criterion to all candidate models and simply choose the best. Physico-chemical criteria can be taken along, by applying model adequacy test as described in section 3.1.

Optimization, by postulating all possible models and finding the global optimum among those, has the advantage that it connects to models based on superstructures. An example is the description of reaction networks. All possible reactions can be suggested or automatically generated. These are incorporated in an overall model that is to fit the available data. The variables are typically the reaction constants expressed as pre-exponential factors and reaction energies. Added to this are binary variables that determine the absence or presence of certain reactions. Physico-chemical criteria can be taken along by adding constraints or weight functions.

Such a problem is therefore restated as an MINLP optimization problem. For example, a model is given with p parameters and $p_1 < p$ parameters are multiplied by a binary variables z_k, $k = 1 \dots p_1$, to account for the presence, $z_k = 1$, or absence, $z_k = 0$, of phenomenon k characterized by parameter β_k. We specify e.g. the error variance of eq. (2) as objective. The optimization problem is then

$$\min_{\beta,\mathbf{z}} \frac{\sum_i \left(\frac{Y_i - f(\mathbf{x}_i;\beta,\mathbf{z})}{\sigma_i}\right)^2}{n - \sum_k z_k}. \tag{6}$$

Petzold and Zhu [6] reported such an approach for chemical kinetics problems. Their objective function was the simple L_2-norm. They reformulated the problem with extra non-linear constraints, which can be interpreted as penalty functions to constrain the sequential optimization process. Also, Edwards et al. [7] heuristically defined additional equality and inequality constraints. Both groups focus on methods to compute the MINLP solution, which is not trivial. Edwards et al. [7] noted that sofar it is proven that the optimization approach is viable, and that it can only improve with the introduction of evolving MINLP algorithms. In the choice of objective function the purpose of the model can be somewhat expressed. As mentioned before, an FPE and AIC (eq. (2)) emphasize either simplicity or complexity.

Model selection in the optimization approach leads to a single candidate model.

5 Bayesian approach

In statistics each event has a probability associated. Such an event is the assignment of a model to a phenomenon, i.e. model selection. Assume a finite set of models with index $m = 1 \dots M$ and initial information, I. A probability, $p(m|I)$, is assigned to each of the events that model m is the valid model, such that

$$\sum_{m=1}^{M} p(m|I) = 1. \tag{7}$$

Any prior physico-chemical information can thus be quantified in this set of probabilities where the engineer or scientist can incorporate existing knowledge in what is effectively a model probability. In the absence of any prior information, it is assumed that all model probabilities are equal and so $p(m|I) = 1/M$. Syvia [8, chapters 4 and 5] gives a readable description of the considerations.

Any new experiment or new data can then be seen as added information that changes the model probability. It is straight forward to calculate the probability or likelihood, $p(\mathbf{y}_{n+1}|m, I)$, of the realization, $\mathbf{y}_{n+1}$, assuming model m is right. Bayes' rule then allows the evaluation of the model probability $p(m|n+1, I)$ after $n+1$ experiments based on the prior model probability $p(m|n, I)$,

$$p(m|n+1, I) = \frac{p(\mathbf{y}_{n+1}|m, I) \times p(m|n, I)}{\sum_{m=1}^{M} p(\mathbf{y}_{n+1}|m, I) \times p(m|n, I)}, \tag{8}$$

where $p(m|0, I) = p(m|I)$. The summation is necessary as often the likelihood only can be calculated.

So starting with the initial probabilities after all N experiments are finished a posterior probability, $p(m|N, I)$, is evaluated

$$p(m|0, I) \Rightarrow p(m|1, I) \Rightarrow \dots \Rightarrow p(m|N, I).$$

If one of the models, m, has a probability, $p(m|N, I)$ exceeding a predefined level, this is then equivalent to the method of hypothesis testing. The advantage of Bayes' approach is that in the case of models that cannot easily be discriminated, a quantitative measure is given, probability, which is easily interpretable and comparable with other requirements of the model.

6 Experimentation for model selection

Experiments can be designed beforehand in order to improve the model selection process. Firstly, the set of possible experiments should be known. Secondly, a set of models that need discrimination are to be defined, either as the result of an engineering process or as the result of a model selection procedure applied to existing data.

The question is to evaluate each of the experimental designs to best discriminate between the given models. Imperative in the experimental design process for non-linear models is the existence of an estimate of the given parameters, either from prior knowledge and expectation of the parameter values or on the basis of model fits from, for example, previous experiments.

Atkinson and Donev [9, chap. 20] deals with this by postulating that one model with known estimates for the parameters is true, and then the parameters of all other models are determined by least squares minimization. This is a valid procedure to obtain parameter estimates for all models as a basis for the estimated responses. The lack-of-fit sum of squares also allows hypothesis testing as in section 3.

The standard approach is to define a discrimination criterion, D, which describes the distance between the estimated responses, $f(\mathbf{x}_{n+1}; \hat{\beta})$ of all models, and to maximize this with respect to the design variables, $\mathbf{x}_{n+1}$.

$$\max_{\mathbf{x}_{n+1}} D(\mathbf{x}_{n+1}), \tag{9}$$

where $n+1$ indicates the next set point. The search space is limited to the space of possible experiments. The theory can be extended with the same ideas to find more set points simultaneously, but then also the specific experimental circumstances should be accounted for. These circumstances determine whether a single or a set of set points is measured with a single experimental effort.

A relatively simple discrimination criterion ignores the uncertainty in the estimations. The L_a-norm,

$$D_a(\mathbf{x}_{n+1}) = \sum_{m=1}^{M-1} \sum_{k=m+1}^{M} \left| f_m(x_{n+1}; \hat{\beta}_m) - f_k(x_{n+1}; \hat{\beta}_k) \right|^a, \tag{10}$$

allows to find those set points where the models diverge most. It is easily interpretable and used without much difficulty. When $a = 2$, this is called the T-optimal design in the procedure of Atkinson and Donev [9], who give a clear treatment of this subject.

The obvious extension of the criterion above is to use the sum of squares, $a = 2$ and weigh the sum with the available variance,

$$D_2(\mathbf{x}_{n+1}) = \sum_{m=1}^{M-1} \sum_{k=m+1}^{M} \frac{(f_m(x_{n+1}; \hat{\beta}_m) - f_k(x_{n+1}; \hat{\beta}_k))^2}{\sigma^2_{m,k,n+1}}, \tag{11}$$

where the denominator accounts for the uncertainty in the responses. It comprises the two uncertainties due to the estimates of the two models concerned, and also the measurement variance, σ^2_{n+1},

$$\sigma^2_{m,k,i} = \mathrm{var}(f_m(x_{n+1}; \hat{\beta}_m)) + \mathrm{var}(f_k(x_{n+1}; \hat{\beta}_k)) + \sigma^2_{n+1}. \tag{12}$$

The measurement variance depends sometimes on the response, $f(\mathbf{x}_{n+1}|\hat{\beta}_m)$ or the set point. In the absence of detailed information a constant value can be substituted or it can be neglected.

An extension to the class of divergence criteria above is to account for the model probabilities introduced in section 5.

$$D_B(\mathbf{x}_{n+1}) = \sum_{m=1}^{M} p(m|I) \sum_{k=1, k\neq m}^{M} D_{m,k}(\mathbf{x}_{n+1}), \tag{13}$$

where $D_{m,k}(\mathbf{x}_{n+1})$ is one of the divergence measures specifically evaluated between the responses of model m and k at set point $\mathbf{x}_{n+1}$.

Finally, this summary ends with the entropy measure, which goes back to the work of Reilly [10]. Entropy is defined as $-\sum_m p(m|n, I) \ln(p(m|n, I)$ and this should be reduced with each subsequent experiment. This leads to a divergence criterion:

$$D_S(\mathbf{x}_{n+1}) = \sum_{m=1}^{M} p(m|n, I) \int_y p_m(y) \ln \frac{p_m(y)}{\sum_k p(k|n, I) p_k(y)} dy, \tag{14}$$

where y is the response at the next set point and $p_m(y)$ its probability density function given model m.

Burke et al. [11] compared a few divergence criteria on copolymerization experiments and they concluded that a measure similar to eq. (11) proved effective closely followed by the entropy measure of eq. (14). It is estimated that the number of experiments could be reduced by a factor of two. Recently, Singh [12] developed a more general divergence measure for multi-response data, in order to take into account correlation between responses.

7 Examples and practical considerations

7.1 Ammonia

An extensive example of the application scheme in Figure 1 is given through a study of isotherm models for the adsorption of ammonia by Helminen et al. [13]. They considered 16 different isotherm models and three different 5 different sorbents. Here we will focus on the results of zeolite 13X [13, Tables 4 and 5], where the Henry model was inadequate.

The basic model gives the amount of adsorption, q, as a function of pressure, p, and temperature, T,

$$q(p,T) = q_s \frac{(bp)^a}{1+(bp)^a} \text{ with } b = b_0 \exp\left(\frac{-\Delta H}{RT}\right). \tag{15}$$

This actually presents the Langmuir-Freundlich (LF) form and the Langmuir (L) form. In the latter case the power, a, equals unity, $a = 1$. Secondly, both forms have five different temperature dependencies of the saturation adsorption, q_s, which equals,

$$q_{s0},\ a_1 + \frac{a_2}{T},\ b_1 \exp\left(-\frac{b_2}{T}\right),\ q_{s0}(1-\alpha T) \text{ and } q_{s0}\exp(-\alpha T), \tag{16}$$

respectively. These models are named $LF1, \ldots, LF5$. Next to these 10 models there are 2 based on the Margules theories (VS-M1 and 2), and one each on the Wilson theory (VS-W), the Flory-Higgins expression (VS-FH), and the Dubinin-Astakhov isotherm (DA).

There were no data given to verify model adequacy in each individual case. For example, an estimate of the experimental accuracy was missing. The second step is model reduction with the F-test. In each case, the more complex Langmuir-Freundlich model could not be rejected in favor of the simpler Langmuir model. Also the more complicated Margules model could not be rejected with the approach taken in section 3.2. The results are summarized in Table 1. The Bartlett-test (eq. (5)) actually shows that all models except for the four Langmuir-Freundlich models can be rejected. Strictly, following the argument of section 3 the end here is with four undistinguishable models where even a further ranking has no influence.

Here, it is necessary to reconsider the four remaining models. In fact, they only differ because the temperature dependence is weak and each has a slightly different approach representing the dependence by a straight line, an exponential

Table 1
Summary of fit results after model reduction with nested models and 32 data points. Basic data are from Helminen et al. [13]

Model	No of parameters	s^2_{res}	FPE	AIC	SDD
LF2	5	0.0361	0.0528	-96.29	-85.49
LF3	5	0.0361	0.0528	-96.29	-85.49
LF5	5	0.0361	0.0528	-96.29	-85.49
LF4	5	0.0392	0.0573	-93.65	-82.85
DA	3	0.0924	0.1188	-70.21	-62.34
VS-W	7	0.2016	0.3450	-36.40	-22.67
VS-FH	5	0.2070	0.2946	-41.25	-30.45
VS-M2	5	0.3612	0.5279	-22.59	-11.79

decrease, etc. A physico-chemical argument could be that the describing function for q_s should always have a finite positive value. This leaves only $LF2$, where $q_s = a_0 + a_1/T$ and $a_0, a_1 > 0$, as the desired model. [2]

7.2 Reaction lumping scheme

[14] reported the results of experiments on CHO-reactions taken in a reactor. It concerned decalin cracking, which could be described with a three-reaction model or a four-reaction model. The data were the actual GLC-spectra. The problem was that the different reaction schemes that were proposed, were based on different lumping schemes. Depending on the lumping scheme different species were taken together in one pseudo-component. This meant that, depending on the model, in the GLC-spectra different areas were taken together. This is a rather unusual situation. As the areas were input to the fitting process, each of the different models had different data, but these were intrinsically from obtained from the same experimental data. The straightforward application of the decision tree (Figure 1) could not be done.

Actually, a naive analysis assuming constant errors was first attempted which led to ambiguity. The introduction of a sound error analysis of the areas in the GLC-spectra effectively solved the model selection problem. Besides a scale factor, the same absolute and relative errors could be applied to all 36 or 45 data. The three-reaction model was then heavily favored, as $SS_{\text{res}}/(n_3 -$

[2] Helminen et al. [13] preferred the model $LF5$. They considered mostly the coefficient of regression.

$p_3) = 14$ against $SS_{\text{res}}/(n_4 - p_4) = 113$ for the four-reaction model. From a purely statistical point of view, the weighted sum-of-squares indicates that both models should be rejected in the model adequacy stage. Nevertheless, the more pragmatic practitioners accept these misfits as reasonable, especially as the lumping approach might lead to large errors.

7.3 Error model and model error

A recurring problem in chemical engineering is the tradition that the experimental errors are not well determined from the measurement systems. So, the errors intrinsically assumed in eqs. (1) and (12) is unknown. Secondly, nearly all models are approximations of the reality and also the size of this approximation is unknown. The resulting sum-of-squares, SS_{res}, is a superposition of a pure error component, SS_{pe}, and a model error component, SS_{me},

$$SS_{\text{res}} = SS_{pe} + SS_{me}.$$

There is a small possibility to investigate this. If all measurements are n times replicated, the error component should decrease with $1/n$. In some circumstances measurements are indeed repeated.

As illustration, this is done here with a repeated simulation of the case given in Table 1 for the three models LF3, LF4, and LF5. Model $LF2$ is the base model for the Monte Carlo simulation. The result (Figuur 3) shows that the limiting sum-of-squares, $1/n \to 0$, leads for the three models to a model variance of respectively, $7\%, 75\%$ and 50% of the total variance.

Strictly, speaking Figure 3 shows that ideally speaking the model error is significant for models LF4 and LF5, while for models LF2 and LF3 the measurement variance dominates. Ideally, in the first case discrimination measures should be based on the simple L_a norm, such as in eq. (10). In the latter case, the full eqs. (11) and (12) can be applied. Of course, the choice will be either one of the two discrimination criteria for the total selection problem. In that case, it is advisable to take at least a rough estimate for the model error, as this slightly improves the selection procedure.

7.4 Time series and multi-variable data

Model selection is a major activity in time series analysis as a support for model-predictive control [3]. The problem here is exactly as stated in the introduction. Namely, given the data of a time-series, determine the regression

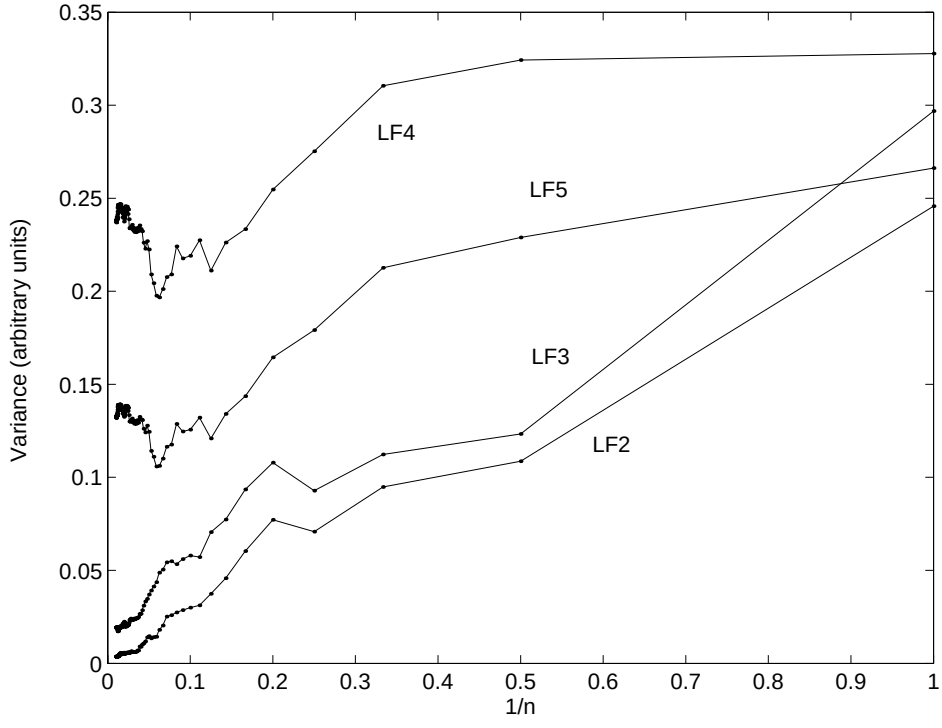

Fig. 3. Illustration of variance reduction with replicated measurements. n is the number of replications.

model that describes the data and that can be used for control purposes. Inputs are, typically, the autocorrelation functions of the signals. Outputs are black-box models, which are normally a severe but adequate simplification of reality. The criteria of eq. (2) are guidelines.

A further step is to consider cross-correlations between various time series. The associated cross-correlation matrices contain the information for principal components analysis, PCA [15]. Standard matrix decomposition techniques are the basis for a dimension reduction, which helps to focus on the important aspects of a model for the whole system. This separation between components describing the core of the model and insignificant components can also be formulated as an inference problem for model reduction. An extension is the consideration of the sensitivity matrices of process systems. The same decomposition technique can be applied namely. Rankin and McCormick [16] give an example of an application in the field of reaction kinetics.

7.5 Ammonia: experimental design

The ammonia data discussed in the first part of this section provides an opportunity to illustrate the design of experiment for model selection. The parameters of model LF2 [13, Table 3] were used as the "correct" model. The

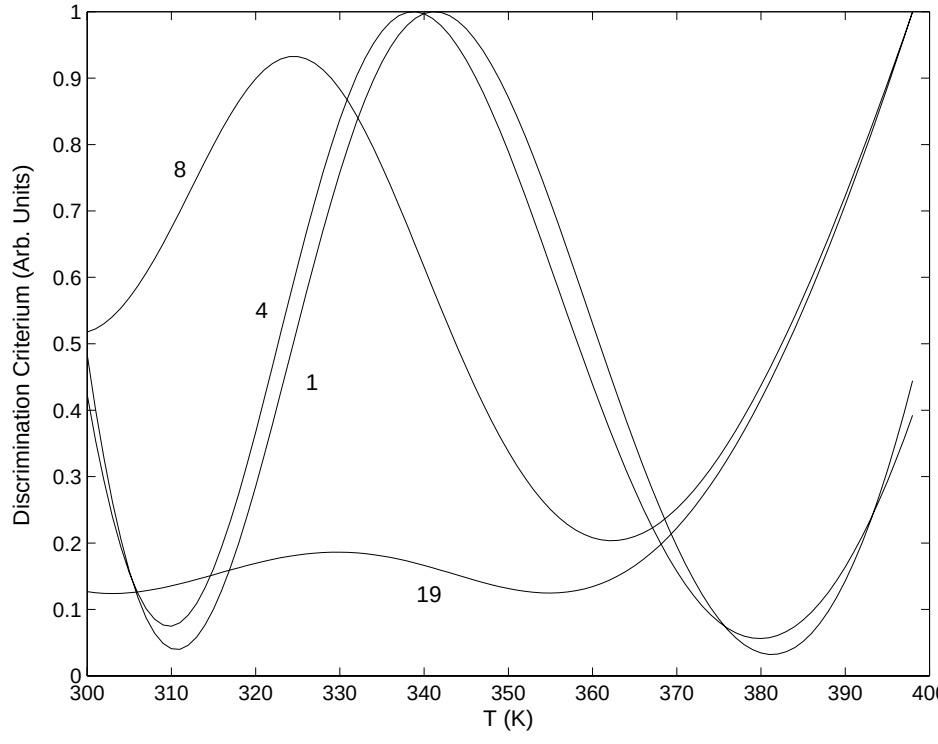

Fig. 4. Discrimination criterion as function of measurement temperature. The numbers near the curves refer to the experiment number in Figure 5.

assumed experimental error was 0.1. An experiment consists of measuring at one temperature, between 298 K and 398 K, and at one fixed series of eight pressures that were directly taken from the original paper [3] .The start point was the simulation of four measurements at four different temperatures. The purpose was to determine the sequence of temperatures for experimentation in order to optimally discriminate between the four independent candidate models, LF2, LF3, LF4 and LF5, of eq. (16) and Table 1. Figure 4 shows the discrimination criterion mentioned in eq. (11) as function of the temperature after 1, 4, 8 and 19 extra experiments. Obviously, according to eq. (9), at first one should measure in the mid-range and at the end at the highest possible temperature. Thus following this recipe, the sequence depicted in Figure 5 was obtained. The Bartlett test statistic (eq. 5) is used here. It namely gives also a clear decision about when to remove a model from the allowable set. Here, after 3 extra experiments LF4 is eliminated, after 7 experiments LF5, and finally LF3 was removed after 19 extra experiments. The sequence and number of experiments needed for reducing the choice to one model was undetermined, but invariably the simulation model was detected as the remaining model.

[3] Contrary to the original paper, four in stead of five parameters were allowed to be estimated. The fifth one, ΔH, would make the system badly ill-conditioned. This can be explained by considering the competing temperature dependencies in the Langmuir-Freundlich formula used.

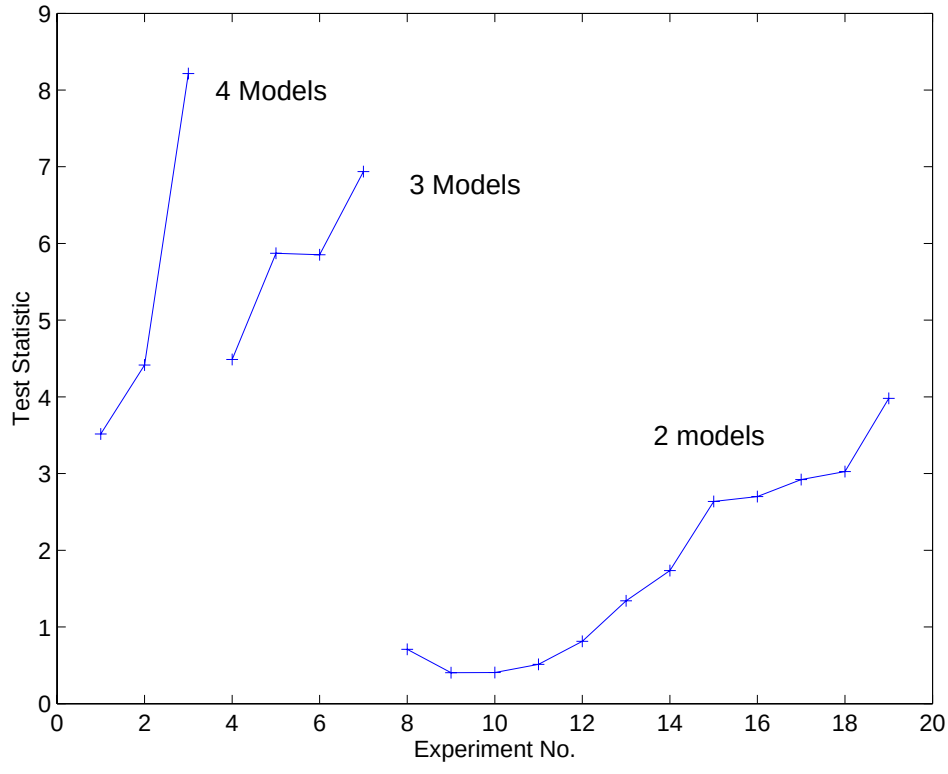

Fig. 5. Test statistic in Bartlett's test, eq. (5) as evaluated after each "experiment".

8 Discussion and conclusion

Model selection is based on the premise that models exist, and that they can be fitted to data or at least be compared with data. An overlooked aspect is that also the error model should be given. This is essential for the inference and Bayesian approach of model selection, as that is mostly based on knowledge of the statistics involved.

However, actual practice is still that models are rough approximations of reality, although the use of more rigorous models in chemical engineering becomes more wide spread. A careful weighing of model error versus experimental error determines the criteria used in especially the optimization-based selection procedures.

The decision scheme (Figure 1) in the inference-based procedure is a good engineering tool, as it translates model selection to a clear-cut procedure. Also the optimization-based selection procedure has this advantage. The first has more possibilities to deals with physico-chemical model requirements, while the latter lends itself more for superstructure-based models [17, chap. 20].

The Bayesian scheme of model selection is best in quantifying the relative

benefits of each of the models. At the end the user has the possibility to make a final selection decision based on knowledge of the probability associated with each model and his expert knowledge of all criteria involved. This is thus more suitable for a scientific environment.

The use of the optimization approach appears still limited to the few references discussed in section 4. Its application might be hampered by the numerical and algorithmic challenges. There are here no cases reported of model selection using Bayes' approach. Reports about its use are dated as the citations by Buzzi-Ferraris [18] and Burke et al. [11] show. In the statistical literature the interest is still continuing in this approach. Chickering and Heckerman [19] suggest some new approaches where also a distinction is made between engineering and scientific criteria.

In normal practice, it is expected that a single model is chosen. As must be clear from above, the inference-based approach may lead to a set of statistically valid models. In the Bayesian approach actually no choice is made, although a user might define a probability level as a lower bound for[13] a valid model. Similarly, the optimization-based selection could use a deviation from the global optimum in order to define a finite set of models. That has its difficulties as each of the objectives used has a specific interpretation, which does not lend itself for defining well argued deviations.

References

[1] D.M. Bates and D.G. Watts. *Non linear regression and its application*. John Wiley, 1988.

[2] H. Linhart and W. Zucchini. *Model Selection*. John Wiley, 1986.

[3] L. Ljung. *System identification: theory for the user*. Prentice Hall, 1987.

[4] D.J. Murray-Smith. Methods for the external validation of continuous system simulation models: a review. *Mathematical and Computer Modelling of Dynamical Systems*, 4:5–31, 1998.

[5] J.A. Romagnoli and M.C. Sánchez. *Data processing and reconciliation for chemical process operatiosn*. Academic Press, 2000.

[6] L. Petzold and W.J. Zhu. Model reduction for chemical kinetics: an optimization approach. *AICHE Journal*, 45:869–886, 1999.

[7] K. Edwards, T.F. Edgar, and V.I. Manousiouthakis. Reaction mechanism simplification using mixed-integer non-linear programming. *Comput. Chem. Eng.*, 24:67–79, 2000.

[8] D.S. Syvia. *Data analysis: a Bayesian tutorial*. Oxford University Press, 1996.

[9] A.C. Atkinson and A.N. Donev. *Optimum experimental designs*. Clarendon Press, 1992.

[10] P.M. Reilly. Statistical methods in model discrimination. *Can. J. Chem. Eng.*, 48:168–173, 1970.

[11] A.L. Burke, T.A. Duever, and A. Penlidis. Discriminating between the terminal and penultimate models using designed experiments: an overview. *Ind. Eng. Chem. Res.*, 36:1016–1035, 1997.

[12] S. Singh. Model selection for a multi-response system. *Chem. Eng. Res. Des., Trans IChemE part A*, 77:138–150, 1999.

[13] J. Helminen, J. Helenius, E. Paatero, and I. Turunen. Comparison of sorbents and isotherm models for nh_3-gas separation by adsorption. *AICHE Journal*, 46:1541–1555, 2000.

[14] M.P. Helmsing. *FCC Catalyst testing in a novel laboratory riser reactor*. PhD thesis, Dept. Chem. Engineering, TU Delft, The Netherlands, 1996.

[15] W.F. Ku, R.H. Storer, and C. Georgakis. Disturbance detection and isolation by dynamic principal component analysis. *Chemom. Intell. Lab. Syst.*, 30:179–196, 1995.

[16] S.E. Rankin and A.V. McCormick. Hydrolysis pseudoequilibrium: challenges and opportunities to sol-gel silicate kinetics. *Chem. Eng. Sci.*, 55: 1955–1967, 2000.

[17] L.T. Biegler, I.E. Grossmann, and A.W. Westerberg. *Systematic methods of chemical process design*. Prentice Hall, 1997.

[18] G. Buzzi-Ferraris. Planning of experiments and kinetic analysis. *Catalysis today*, 52:125–132, 1999.

[19] D.M. Chickering and D. Heckerman. A comparison of scientific and engineering criteria for bayesian model selection. *Stat. Comp.*, 10:55–62, 2000.