Handling Uncertainty in the Development and Design of Chemical Processes

David Bogle, David Johnson, and Sujan Balendra
Dept of Chemical Engineering, University College London,Torrington Place, London, U.K. WC1E 7JE
e-mail: d.bogle@ucl.ac.uk

Abstract. The paper presents a stochastic methodology for handling uncertainty in process development as part of a general framework for batch and continuous process models. The method combines systematic modelling procedures with Hammersley sampling based uncertainty analysis and a range of sample-based sensitivity analysis techniques, used to quantify predicted performance uncertainty and identify key uncertainty contributions. The methodology was implemented on a batch reactor process and some clear recommendations as to how to reduce the uncertainty in the main output variables are obtained. The paper concludes with some discussion about an alternative approach to use instead error bars from experimental data as intervals and using interval methods to determine the best 'worst-case' design.

1. Introduction

In the development of new chemical manufacturing processes, particularly in the pharmaceutical industry, there is a large element of process uncertainty since detailed knowledge of the chemical reaction mechanisms and of the power and effectiveness of separation devices (to purify the product and recover raw materials) is often limited. Data is obtained from the laboratory during the identification stage of a new product (for example a new drug) and this is used during the manufacturing process development stage. Much data is generated but often not useful for the development of the large scale manufacturing process. In some cases large amounts of data are available but often single data points with confidence limits are obtainable in the form of interval bounds. Using a structured approach with the computational process design tools, which are used extensively, the uncertainty can be managed and improved process performance may be obtained. The methodology proposed is based on a stochastic formulation but the use of interval methods which have a natural role arising from the form of data used is also discussed briefly.

This work was undertaken for process development in the pharmaceutical industry. New products are constantly being proposed and the decision about whether to proceed with development depends on the efficacy of the drug but also whether the manufacturing process will be possible and will make a profit. It is therefore necessary to develop new processes but this is often done without regard to the data being obtained in the laboratory and without consideration of the accuracy of that data. The objective was to develop a model based approach that could help identify major causes of uncertainty and hence help to direct when better data is required. Much good data is developed but often the data required for manufacturing process development is not available or of poor quality.

Pharmaceutical processes typically consist of a sequence of unit operations, for example reactors and separation devices. It is important that the methodology is able to handle a large sequence of units as well as single units. The main causes of uncertainty in process development are in the data obtained about reaction and separation which are then used in the model and also in the quality of the raw materials that are used in the reactions. Assumptions about models are also uncertain which can cause the mathematical structure to be incorrect as well as the model parameters, for example in the case of the order of the reaction kinetics.

While the approach was developed with pharmaceutical processes in mind it could in fact be used for any type of process.

2. A Methodology for Design Under Uncertainty: Combined risk analysis and systematic model development

The proposed methodology aims to introduce some form of management of the uncertainty associated with the model representations of the current process knowledge. It is assumed that the conceptual process design (equipment allocation and design) is already decided. The deterministic process models may exhibit non-linear and dynamic characteristics as may be expected in typical pharmaceutical processes. However, spatially distributed models are not considered for computational reasons.

In the face of large amounts of uncertainty predicted in the important process output criteria, three issues have been considered:

(i) reduction of the uncertainty by improving current models/parameter estimations associated with the key contributing uncertainty factors identified,
(ii) manipulation of the available process decisions (operating policy) to improve process robustness to model uncertainty,
(iii) consideration of process alternatives.

Issue (i) concerns the gathering of additional information for systematic model development for more reliable models. Issues (ii) and (iii) concern the optimisation and comparison of uncertain but integrated processes sequences. This will be dealt with in a future paper.

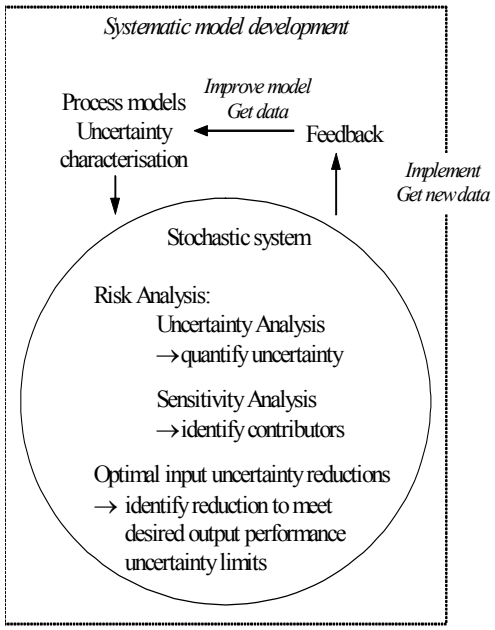

Figure 1. Management of uncertainty in a model-based approach to integrated design under uncertainty for pharmaceutical processes.

The elements of the Risk Analysis approach have been combined with systematic procedures for the development of deterministic process models (Figure 1). A stochastic representation of the complete model of the integrated process sequence is generated to quantify modelling uncertainties and to identify and rank the most important contributors in the uncertain (but

structured) system with respect to the important system responses. The suggestion is that this information can be used to drive the general direction for data collection (within process development) to improve the key models and reduce the uncertainty in the most significant areas. As more data becomes available the methodology allows the tracking of the effect of increased knowledge in certain process models and the effect this has on the complete system, in an iterative manner. A key issue is the flexibility of the approach to incorporate new data into the analyses. A more detailed schematic of the approach is shown in Figure 2. The reader is referred to Hangos and Cameron (2001) for further detail about conventional model development.

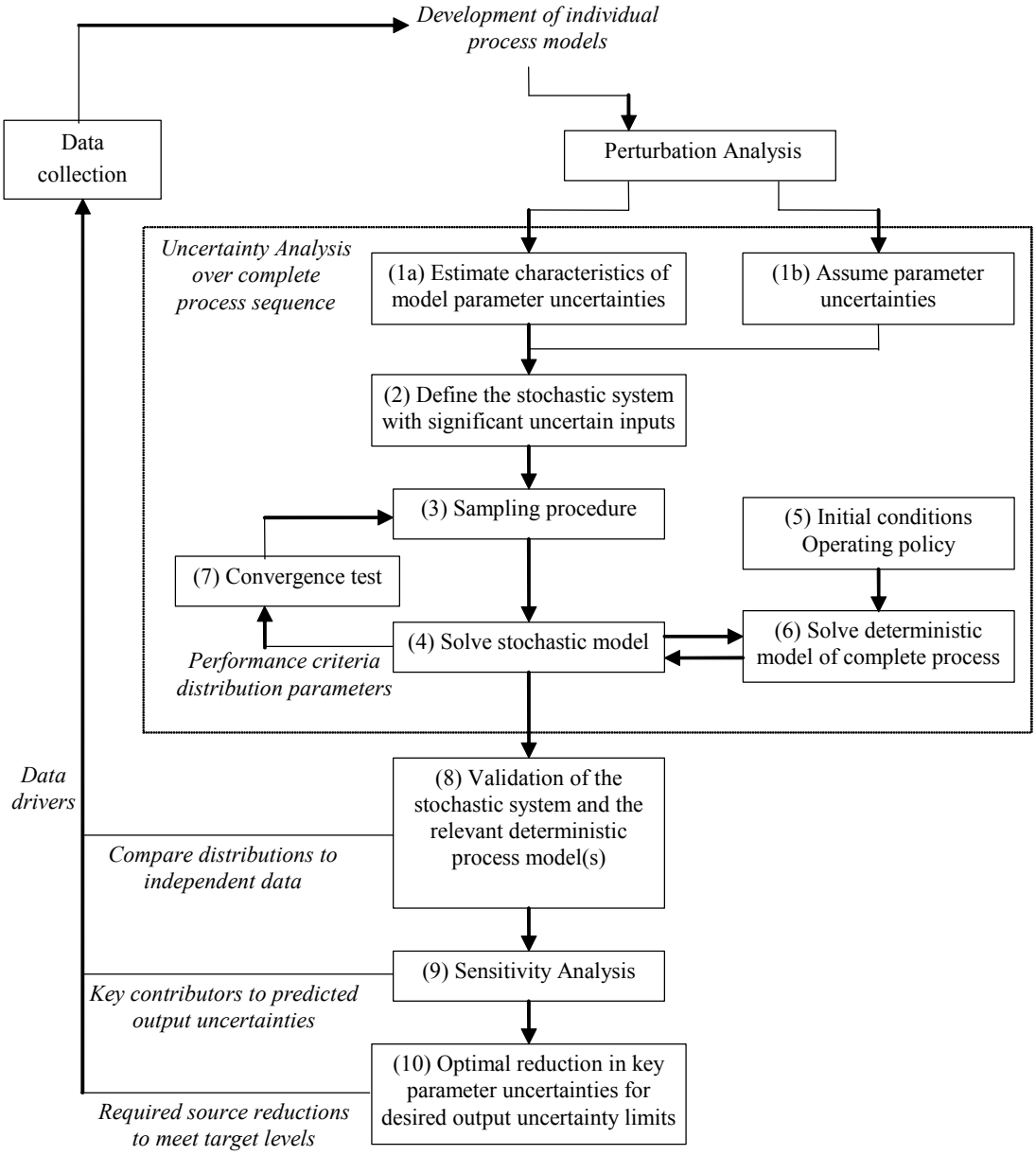

Figure 2. Schematic for the systematic model development incorporating the Risk Analysis approach under uncertainty.

Once a model has been developed and before the Risk Analysis methods are implemented a screening procedure is used to determine which of the parameter uncertainties in the complete

sequence model may be potentially significant regarding the response variables in the stochastic representation of the system. This is necessary when the number of parameters increases and it is necessary to limit the number of dimensions in the following Risk Analysis. For this purpose a Perturbation Analysis (one at a time approach) is used since it is a common and easy method to implement. The deterministic model is systematically simulated at positive and negative deviations from the nominal parameter estimates. The magnitude of the deviations may be based on the judgement of the developer (for example these could be the approximate precision ranges for different types of model parameters as suggested by Hangos and Cameron, 2001), or at estimated confidence limits if available.

3. Uncertainty Analysis

Following the selection of the significant uncertain parameters, Steps 1 to 7 in figure 2 comprise the elements of Uncertainty Analysis. The information required for the stochastic distributions of the input variables needs to be developed typically from sparse data sets so the methodology includes the development of these distributions. In Step 1 the approach used for the quantitative estimation of the uncertainties in these parameters is determined by the data available for parameter estimation which may be based on three different information sources:

- analysis of the performance of the model building based on experimental measurement data (Step 1a), allows the estimation of uncertainty in the parameter estimates using confidence intervals or regions,
- expert technical judgement is needed when quantitative data is not available for systematic model building and models are assumed whose parameter values are instead based on observations and/or assumed along with associated confidence intervals and probability distributions (Step 1b),
- either specific published information or information from which judgements can be inferred (Step 1a or 1b).

If parameters are estimated or assumed independently of each other, the joint sampling space may be described as a hyper-rectangle where each dimension represents one uncertain input bounded by its respective upper and lower confidence limits. If no data is available for model parameter estimation, confidence limits around the nominal values are assumed as some percentage of the nominal. For uncertainty in independent parameters of assumed nominal values, to be characterised by normal distributions, the standard deviation is assumed at some percentage of the nominal value. Confidence limits around the nominal value can be assumed at some number of standard deviations (typically two or three deviations for approximately 95 or 99.9% probability of containment according to Chebyshev's rule).

Least squares regression is a commonly used parameter estimation method for which confidence intervals can be simply stated assuming normally distributed uncertainty. Although likelihood and lack of fit are more accurate methods for estimating parameter confidence regions for non-linear models, Donaldson and Schnabel (1987) conclude from their general study on regression parameter confidence regions that the linearization methods provide the most concise representation of information required to construct confidence intervals and regions.

For a model that is non-linear in its parameters, individual confidence intervals can be approximated assuming a linearization of the model about its optimal estimated parameter values, θ',

$$\left|\theta_p - \theta'_p\right| \le \left(\hat{V}_{pp}\right)^{\frac{1}{2}} t_{N-P,1-\frac{\alpha}{2}} \tag{1}$$

where subscript p is the index of the input uncertainty (θ), V is the covariance matrix, and the values of the confidence limits are defined where the value of t is taken from the Student's t-test

distribution with N-P degrees of freedom (number of regression data points, N, and number of parameters, P, in the regression model), assuming a desired level of confidence, 1-α. In a multi-parameter model where the parameters are estimated simultaneously, a joint confidence region provides a more appropriate measure of the (normally distributed) uncertainty space that would be a hyper-rectangle comprising the individual confidence intervals. Similarly, for a non-linear model, a hyper-ellipsoidal joint confidence region is approximated by,

$$\left(\theta-\theta'\right)^T \hat{V}^{-1}\left(\theta-\theta'\right) \le PF_{N,N-P,1-\alpha} \tag{2}$$

where θ is a vector of the model parameters and the value of F is taken from the F distribution. This is the distribution of a random variable, F, defined as the ratio of two independent chi-squared random variables divided by their respective degrees of freedom.
Linearization methods for the estimation of confidence intervals and regions require the estimation of the parameter covariance matrix. Donaldson and Schnabel (1987) state that the most common and easily computed estimate for the covariance matrix is,

$$\hat{V} = s^2\left(J\left(\theta'\right)^T J'\right)^{-1} \tag{3}$$

where $J(\theta')$ is the Jacobian matrix of the model predictions at the optimal parameter estimates (i.e. the N × P matrix with the (n, p)th element estimated by $\partial f(x_n, \theta)/\partial\theta_p$ at θ′, for N data points and P parameters), s^2 is the estimated residual variance computed from the residual sum of squares (RSS) between the regression model predictions, $\hat{\Phi}$, at the optimum parameter estimates and the measurement data, Φ,

$$s^2 = \frac{RSS\left(\theta'\right)}{\left(N-P\right)} \tag{4}$$

$$RSS\left(\theta'\right) = \sum_{n=1}^{N}\left(\Phi_n - \hat{\Phi}_n\left(\theta'\right)\right)^2 \tag{5}$$

and where $\hat{V}_{pp}$ is the ppth element of the covariance matrix, $\hat{V}$, and is the variance estimate of the pth model parameter (input uncertainty). J is estimated numerically using the first order Taylor's approximation J by introducing deviations into each optimal parameter value in turn and re-evaluating the change in the predicted dependent variable at each data point.
Given the covariance matrix it is straightforward to determine the correlation matrix, $\hat{C}$,

$$\hat{V} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1P}\sigma_1\sigma_P \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2P}\sigma_2\sigma_P \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{P1}\sigma_P\sigma_1 & \rho_{P2}\sigma_P\sigma_2 & \cdots & \sigma_P^2 \end{bmatrix} \quad \Rightarrow \quad \hat{C} = \begin{bmatrix} 1.0 & \rho_{12} & \cdots & \rho_{1P} \\ \rho_{21} & 1.0 & \cdots & \rho_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{P1} & \rho_{P2} & \cdots & 1.0 \end{bmatrix} \tag{6}$$

where ρ is the correlation coefficient and σ is the parameter standard deviation (determined from the square root of the parameter variances, σ^2, in the leading diagonal of the covariance matrix).

Step 2 defines the stochastic system considered by combining the deterministic process model sequence with the uncertain parameter characterisations as obtained in step 1.

A sampling procedure is invoked in Step 3 in order to approximate the uncertain system. In this methodology, the quasi-Monte Carlo Hammersley Sequence Sampling (HSS, Diwekar and Kalagnanam, 1997) is implemented. Sampling approaches are the most flexible because of their capacity to capture different perspectives of risk, the examination of the entire Θ-space, and they are not severely limited by the number of dimensions, of which HSS appears to be the most efficient. A unit hyper-rectangle of dimension P is sampled using HSS in Step 3.
Diwekar and Kalagnanam (1997) define the M points of the Hammersley sequence variant, $e_p(m)$ in a P-dimensional hyper-cube. Rank correlation coefficients are a meaningful way to describe dependencies between stochastic inputs. The desired rank correlation matrix, $\hat{C}^*$, of a matrix of independently generated sample input column vectors, X, is set as equal to $\hat{C}$ (the desired correlation matrix of X). A new matrix, K, is defined which has the same dimension as X but is independent of X, to give a correlation matrix close to the identity matrix. These are inverted over the standard normal cumulative distribution and the elements in K are rearranged to obtain the correlation structure defined by $\hat{C}$, to give a matrix, K^*. Not only is it necessary that the correlation matrix of K is close to the identity matrix but also that the correlation and rank correlation matrices of K^* should be close to each other.

The stochastic system is solved in Step 4, to obtain probability distributions and distribution parameters for the desired process performance variables (i.e. the desired output variables). This is achieved by sequential simulation of the deterministic model in Step 6 at each observation of the uncertain parameters and at the initial conditions and operating conditions fixed in Step 5, given the matrix of stochastic input observations with any induced correlation structures (X^*) and the deterministic model of the complete flowsheet. To terminate the solution of the stochastic model, a convergence test is employed (Step 7). The convergence test used in this methodology is a tolerance limit on the average sum of squared deviations measured in a distribution parameter, w, over all previous iterations up to the current iteration observation, m_i. This limit, Δw, for the qth predicted process output quality criterion is shown in Equation 7,

$$\Delta w_q = \frac{1}{m_i}\sqrt{\sum_{m=1}^{m_i}\left(w_{q,m} - w_{q,m_i}\right)^2} \le \varepsilon_{w,q} \qquad (7)$$

where w is the mean or variance estimate from all the previous observations and ε is a permitted tolerance. The test requires that tolerances on the mean and variance parameters characterising the distributions in the key outputs are met.

4. Validation

The individual models of the process sequence need to be validated with available independent data of good quality. In the case of using data from a pilot plant run, data for individual operations may not be available since measurements are not taken at all points in the sequence. Here, validation may only be possible over sub-sequences of integrated models. In Step 8, a form of statistical model validation compares distributions of performance predicted from Uncertainty Analysis with independent data to validate the uncertain sequence model. Both the location and spread of the predicted distributions in relation to the independent data are important in the validation. Independent data may already be available from previous runs or if resources permit from specific model validation runs (for specific operations). As stated by Basu et al. (1999)

there should be plenty of opportunities to obtain more data for this purpose given the nature of pharmaceutical process development.

Following this, Sensitivity Analysis is used to estimate the ranking priority of the key stochastic inputs contributing to the uncertainty in the stochastic process output criteria (Step 9). In an efficient manner the sensitivity techniques employed in this methodology reuse the sample results generated from Uncertainty Analysis to avoid the need for any further simulations of the deterministic model.

Standardised regression coefficients (SRC) may be compared to correlation coefficients (CC) to avoid the affect of spurious correlations to which the CCs are susceptible. SRCs may be calculated either from first determining the linear regression coefficients, b_p, and then multiplying these by the parameter sample standard deviation, s_θ, and dividing by the output standard deviation, s_Φ,

$$SRC_p = \frac{b_p s_{\theta_p}}{s_\Phi} \tag{8}$$

or by standardising the raw sample data and then applying the regression,

$$\theta_{p,std,m} = \frac{\theta_{p,m} - \overline{\theta}_p}{s_{\theta_p}} \quad , \quad \hat{\Phi}_{std,m} = \frac{\hat{\Phi}_m - \overline{\Phi}}{s_\Phi}$$

$$\hat{\Phi}_{std,m} = \sum_{p=1}^{P} SRC_p \theta_{p,std,m} \tag{9}$$

where $\overline{\theta}$ and $\overline{\Phi}$ are the sample means of θ and $\hat{\Phi}$, and the subscript 'std' represents a standardised value. To avoid over-fitting problems in determining SRCs, stepwise regression procedures are employed.

The input sample set is split into a number of disjoint intervals which each contain an equal number of observations. In this way the conditional means of the outputs at given values of the inputs can be approximated for the first order CR_p for θ_p,

$$CR_p^2 = \frac{Var_{\theta_p}\left\{E\left\{\Phi\middle|\theta_p\right\}\right\}}{Var\left\{\Phi\right\}} = 1 - \frac{E_{\theta_p}\left\{Var\left\{\Phi\middle|\theta_p\right\}\right\}}{Var\left\{\Phi\right\}} \tag{10}$$

where Φ is the vector of deterministic model performance outputs, θ_p is the vector of observations in the pth uncertain input, Var_{θ_p} and E_{θ_p} are the variance and expectation condition for θ_p. If there is any element of doubt then scatter plots between individual input-output pairs can be viewed, though these may also be susceptible to spurious correlations.

Following identification of the critical uncertain parameters from Sensitivity Analysis, the methodology provides the possibility to determine the minimum reduction in these uncertainties required to meet desired levels of reduction in the uncertainty contained in the performance criteria of the original system (Step 10). A quantitative estimate of the minimum extent of reductions required in the important uncertainty sources to meet desired output uncertainty levels can be provided by formulating a stochastic optimisation problem. In addition, trade-off curves between different parameter reductions can be plotted by solving at different levels of desired performance uncertainty reduction.

By defining the decision variables as the fractions of the original values (before uncertainty reduction) of the parameters which characterise the spread of the parameter uncertainties and formulating the objective function, $\Im$, as a summed term of these decisions, the desired problem formulation is obtained. Since only normal and bounded range parameter uncertainties are currently considered in the combined modelling and Risk Analysis part of the methodology, the

space size characterising parameters, δ, in the optimisation are the standard deviation, σ, for normally distributed uncertainties, p_N, and the deviation of the limits, θ^{UB} and θ^{LB}, from the mean, μ, for bounded range uncertainties, p_U. The values of these decisions are passed to the HSS sampling sub-routine which locates observations within the redefined uncertainty space. The new stochastic model is solved given the fixed initial conditions, operating policy and remaining set of parameters. The objective is maximised subject to inequality constraints which permit a fraction, α, of the original level of the uncertainties observed in the original output variables. A stochastic optimisation algorithm is used to solve the following problem:

$$\max_{\delta_{p_N},\delta_{p_U}} \quad \Im = \sum_{p_N=1}^{P_N} \delta_{p_N} + \sum_{p_U=1}^{P_U} \delta_{p_U} \tag{11}$$

Subject to the deterministic model equations and constraints and

$$FW\left(\Phi_q\right) \le \alpha_q FW\left(\Phi_q\right)' \qquad \forall \quad q = 1\ldots Q$$

$$0 < \delta_{p_N} \le 1 \qquad \forall \quad p_N = 1\ldots P_N$$

$$0 < \delta_{p_U} \le 1 \qquad \forall \quad p_U = 1\ldots P_U$$

$$\Theta_N = \left\{ \theta_{p_N} \middle| N\left(\mu'_{p_N}, \delta_{p_N}\sigma'_{p_N}\right) \right\} \qquad \forall \quad p_N = 1\ldots P_N$$

$$\Theta_U = \left\{ \theta_{p_U} \middle| U\left(\mu'_{p_U} - \delta_{p_U}\Delta\theta_{p_U}, \mu'_{p_U} + \delta_{p_U}\Delta\theta_{p_U}\right) \right\} \qquad \forall \quad p_U = 1\ldots P_U$$

$$\Delta\theta_{p_U} = \theta_{p_U}^{UB'} - \mu'_{p_U} = \mu'_{p_U} - \theta_{p_U}^{LB'} \qquad \forall \quad p_U = 1\ldots P_U$$

where the indices o, s, m, d and q are associated with the initial conditions, process stages, parameter scenarios, operating policy variables (z) and performance criteria (Φ). The time invariant policy variables, υ, the time dependent variables, ν, and the stage duration times, t_f, remain fixed at the values specified in the prior Risk Analysis. The measure of uncertainty in the performance criterion, Φ, is the width between the 5 and 95% fractiles, FW(Φ). The prime represents the original value before uncertainty reduction. The total uncertainty space, Θ, is the combined space of the normal and uniform spaces Θ_N and Θ_U. The Matlab Sequential Quadratic Programming routine was used to solve these problems.

It is assumed that the original values of the distribution means (nominal parameter values) are maintained. If the stochastic problem contains decisions in linearly correlated inputs, it is assumed that a change in the spread of one of the correlated parameters gives an equivalent change in the others, while maintaining the same correlation structure.

The solutions of these problems can provide a quantitative idea of the required efforts in reducing specific parameter uncertainties compared to returns in performance uncertainties, which may be used to support data collection decisions.

This information, combined with that obtained from model validation (Step 8) and Sensitivity Analysis (Step 9), provides a useful breakdown of the information required to focus relative experiment planning and data collection efforts towards improving a specific process model within the sequence (by inferring the key uncertain phenomena associated with the identified process sub-sequence and parameter uncertainties), with respect to the possible relative benefits which may be obtained in doing so. The data driver feedback loop shows the position of experiment planning and data collection in Figure 2, though specific decisions regarding these procedures are not explicitly considered in this work.

5. Case study: A multiphase semi-batch reactor process

This case study is based on an exothermic multiphase reaction process and kinetic model investigated by Sano et al. (1998). This case study is a single unit operation but the methodology has been used on a sequence of unit operations (Johnson). The process is for the production of a pharmaceutical intermediate, formed from the amination of a bromopropyl compound. Sano et al developed a kinetic model based on reaction calorimetry data obtained under laboratory conditions in order to determine the optimum feasible and safe operating policy. There is considerable uncertainty in many of the experimental parameters and even in the assumptions underlying the model.

Solid particles of the active pharmaceutical ingredient (API) bromopropyl feed compound (A) reside in an organic solvent (methanol) inside the reaction vessel. A fixed volume of a 50 wt% aqueous dimethylamine reagent (B) is added to the vessel at a constant flowrate under continuous agitation. The solids gradually dissolve and react with the dimethylamine. A diagram of the process is shown in Figure 3. The exact physico-chemical phenomena for this process are not known. The reaction consists of a parallel-series reaction in which the dimethylamine reacts with the dissolved API feed to form the desired intermediate (C) which in turn reacts with the active feed (A) to form a dimeric byproduct (D) in parallel,

$A + B \rightarrow C$	Main reaction
$A + C \rightarrow D$	Sub-reaction

By-product D is known to be very difficult to remove in the downstream purification stages. Intrinsic first order reaction kinetics are assumed in the deterministic process model proposed by Sano et al. (1998) but this is a source of uncertainty. An initial rate limiting period due to the dissolution of solids B, was observed to be independent of solvent concentration and agitation speed within the range of conditions approved. A crude approximation of first order kinetics (with Arrhenius constant and activation energy) is assumed in the model for this dissolution controlled period. This period was observed to last until approximately 55% conversion of A for all the conditions considered, at which point the reaction appeared to be limited by the intrinsic reaction kinetics.

The kinetic model is combined with a standard semi-batch reactor model with constant volume addition (of reagent B). The model equations are given in the Appendix. Consideration of the cooling capacity of the reactor resulted in a limiting relationship between the operating policy variables of feed B addition time, t_{add}, and isothermal temperature, T_{iso}. For the purposes of this study, this relationship is well approximated with T_{iso} as a quadratic function of t_{add} since data regarding the energy balance is unavailable, where the nominal values of the constants C_1, C_2 and C_3 are 7.06, -43.50 and 352.67 respectively.

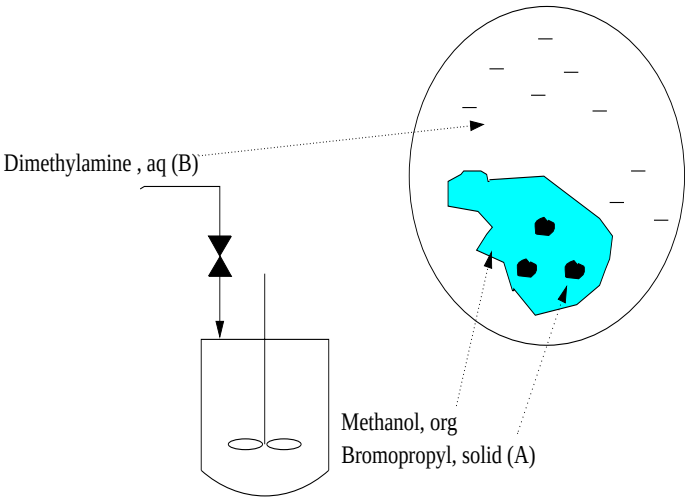

Figure 3. Multiphase batch reaction process

One of the process development objectives for which the model would be used is to help determine the best operating conditions for maximum product yield, Y_C. A reaction time, t_f, of less than 8 hours (terminated when the rate of conversion of A falls below 0.1%) and a final yield in the impurity, Y_D, of below 2% must be maintained. The model was optimised to obtain a nominal set of operating conditions which maximise the yield of C.

Of course uncertainty in the model parameters could have a large effect on any results predicted by the model. This may be of particular importance regarding the optimal operating policy determined subject to the desired limits on process performance. Hence the methodology presented in figure 2 was implemented on the case study.

Perturbation Analysis indicates 11 uncertain parameters which appear to have a non-negligible influence on yield of C (Y_C), yield of D (Y_D) and the final time (t_f): the kinetic rate law parameters ($Ea_{1,int}$, $A_{1,int}$, $Ea_{2,int}$, $A_{2,int}$, $Ea_{1,diss}$, $A_{2,diss}$), the conversion related transition point from dissolution controlled kinetics to intrinsically controlled kinetics, ($X_{A,diss}$), the molar ratio of active feed (m_{A0}) to reagent feed (m_{B0}) and the quadratic constants of the safety constraint (C_1, C_2 and C_3). The assumed uncertainties of these parameters are quantified in Table 1. Correlations are assumed between the activation energy (Ea) and the natural logarithm of the Arrhenius coefficient (A) parameters for each reaction rate constant and between the safety constraint constants.

A total of 490 scenarios were required to satisfy the convergence criterion of 0.5% error in the mean and variance parameters for both Y_D and t_f. The key results of an Uncertainty Analysis under the nominal optimum isothermal operating conditions are shown in Table 2. Under uncertainty in the model parameters the process is predicted to perform particularly poorly regarding violation of the safety constraint for T_{iso}, with an expected probability of passing of only 0.281. The probability of passing the Y_D constraint (at most 2%) is only 0.670 with an expected extent of violation of 0.245%. However, the corresponding values for t_f appear to perform better (at most 8 hours).

Table 1. Uncertainty characterisation in the parameters of Case Study.

	Normal distribution, N(μ, σ, $\hat{C}$)
Ea_1, A_1	$N\left(\begin{matrix} Ea_1^* = 90.46\times10^4 & \sigma_{Ea_1^*} = 10\%Ea_{1s}^* \\ A_1^* = 4.68\times10^{15} & , \sigma_{A_1^*} = 10\%A_1^* \end{matrix}, \hat{C} = \begin{bmatrix} 1.00 & 0.99 \\ 0.99 & 1.00 \end{bmatrix}\right)$
Ea_2, A_2	$N\left(\begin{matrix} Ea_2^* = 65.97\times10^4 & \sigma_{Ea_2^*} = 10\%Ea_2^* \\ A_2^* = 1.00\times10^{10} & , \sigma_{A_2^*} = 10\%A_2^* \end{matrix}, \hat{C} = \begin{bmatrix} 1.00 & 0.99 \\ 0.99 & 1.00 \end{bmatrix}\right)$
$Ea_{1,diss}$, $A_{1,diss}$	$N\left(\begin{matrix} Ea_{1,diss}^* = 78.98\times10^4 & \sigma_{Ea_{1,diss}^*} = 10\%Ea_{1,diss}^* \\ A_{1,diss}^* = 1.76\times10^{13} & , \sigma_{A_{1,diss}^*} = 10\%A_{1,diss}^* \end{matrix} \hat{C} = \begin{bmatrix} 1.00 & 0.99 \\ 0.99 & 1.00 \end{bmatrix}\right)$
$X_{A,diss}$	N(0.55 ± 0.05)
$m_{A0,ratio}$	N(0.357 ± 10% nominal)
C_1, C_2, C_3	$N\left(\begin{matrix} C_1^* = 7.06 & \sigma_{C_1} = 1.15 \\ C_2^* = -43.50 & \sigma_{C_2} = 4.77 \\ C_3^* = 352.67 & \sigma_{C_3} = 4.53 \end{matrix}, \hat{C} = \begin{bmatrix} 1.00 & -0.99 & 0.97 \\ -0.99 & 1.00 & -0.99 \\ 0.97 & -0.99 & 1.00 \end{bmatrix}\right)$

Table 2. Uncertainty Analysis results under nominal optimum isothermal operating conditions

	Y_C (%)	Y_D (%)	t_f (hr)	safety constraint
Expected value	96.34	1.79	6.33	-
Expected extent of constraint violation	-	0.245	0.118	0.440
Probability of passing	-	0.670	0.876	0.281

Sensitivity Analysis (step 9) shows which of the parameters are identified as key to the important output criteria. Correlated parameters are not included in the analysis since the presence of strongly correlated inputs invalidates the linear regression for the standardised regression coefficients. Approximate correlation ratios, Figure 4, show that the variance in the activation energy of the intrinsic parallel reaction, Ea_2 which is parameter number 2 in the figure (and the Arrhenius parameter, A_2, through correlation and the assumption of a linear joint confidence region), is the key uncertain parameter affecting the uncertainty in the prediction for both Yield D and Yield C. No single uncertain parameter is identified as being the main contributor to the uncertainty in the final time, t_f.

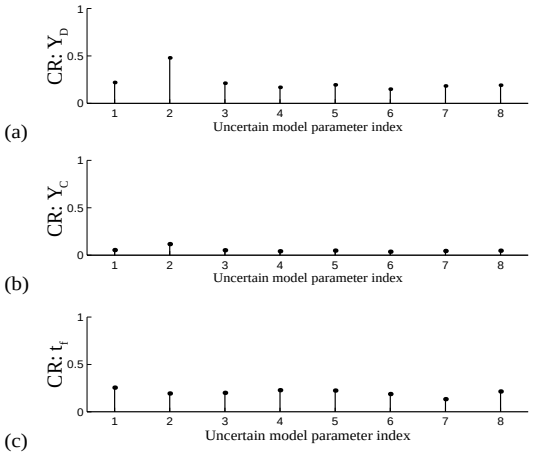

Figure 4. Sensitivity Analysis results under nominal optimum isothermal operating conditions

Further information quantifying the potential uncertainty reduction requirements to meet levels of reduction in the Y_C, Y_D and t_f criteria (step 10) is obtained from the solution of the following problem at different levels of desired output uncertainty reduction:

$$\max_{\delta_{p_N},\delta_{p_U}} \sum_{p_N=1}^{P_N} \delta_{p_N} + \sum_{p_U=1}^{P_U} \delta_{p_U}$$

subject to

deterministic process stage model equations and the following conditions on the 5% and 95%fractile requirements of the products ranges, constraints on δ to ensure the range is not exceeded, and the uncertainty description (θ)

$$FW_{5\%,95\%,Y_C} \le \alpha_{Y_C} FW'_{5\%,95\%,Y_C}$$

$$FW_{5\%,95\%,Y_D} \le \alpha_{Y_D} FW'_{5\%,95\%,Y_D}$$

$$FW_{5\%,95\%,t_f} \le \alpha_{t_f} FW'_{5\%,95\%,t_f}$$

$$0 < \delta_{p_N} \le 1,$$

$$0 < \delta_{p_U} \le 1$$

$$\Theta_{p_U} = \left\{ \theta_{p_U} \middle| U\left(\mu'_{p_U} - \delta_{p_U}\Delta p_U, \mu'_{p_U} + \delta_{p_U}\Delta p_U\right) \right\}$$

$$\Delta p_U = p_U^{UB'} - \mu'_{p_U} = \mu'_{p_U} - p_U^{LB'}$$

$$\Theta_{p_N} = \left\{ \theta_{p_N} \middle| N\left(\mu'_{p_N}, \delta_{p_N}\sigma'_{p_N}\right) \right\}$$

where $p_N = \left\{ Ea_{1,\text{int}}, Ea_{2,\text{int}}, Ea_{1,diss}, A_{1,\text{int}}, A_{2,\text{int}}, A_{1,diss}, C_1, C_2, C_3 \right\}$

$$p_U = \left\{ m_{A0,ratio}, X_{A,diss} \right\}$$

The stochastic optimisation in step 10 determines the minimum reduction in the uncertain parameters, θ_{p_N} and θ_{p_U} required in order to maintain stochastic inequality constraints for α% reductions in the widths of the predicted 5-95% fractile intervals for Y_C, Y_D and t_f from their original values. The parameters manipulated are the fractions of the original standard deviations of the normally distributed uncertain parameters, $Ea_{1,\text{int}}$, $Ea_{2,\text{int}}$, $Ea_{1,\text{diss}}$, C_1, C_2, C_3 and those of the

original bounding widths about the means of the uniformly distributed uncertain parameters, $m_{A0,ratio}$ and $X_{A,diss}$. Equivalent reductions in the uncertainty in the correlated parameters ($A_{1,int}$, $A_{2,int}$, $A_{1,diss}$) are assumed, to maintain the original correlation structures.

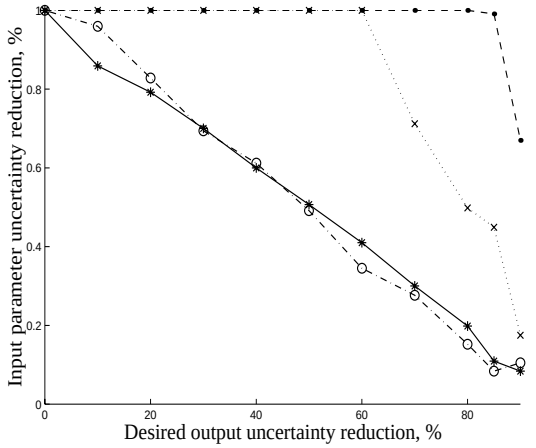

Figure 5. Model parameter uncertainty reductions meeting desired output criteria uncertainty reduction levels.
Key: ∗— = $\sigma_{Ea1,int}$, ×···· = $\sigma_{Ea1,diss}$, •--- = $\Delta m_{A0,ratio}$, o·-· = $\sigma_{Ea2,int}$.

The key results of these optimisation problems are shown in Figure 5. Reduction of the uncertainty in the intrinsic product and by-product activation energies (and the correlated Arrhenius constants) is shown to reduce the uncertainty in the output criteria to levels of around 60% of the original predicted uncertainties, at a constant rate. As uncertainty in the key input parameters is reduced in order to meet the desired levels of uncertainty in the output criteria, the contributions of the uncertainty in other input parameters become important and additionally need to be reduced. This is indicated in Figure 5 at levels of 60% and 90% output uncertainty reductions, where the respective optimal solutions state that reductions in the uncertainty in the dissolution activation energy (and the corresponding Arrhenius constant) and the API feed ratio become relatively more important than in $Ea_{1,int}$ and $Ea_{2,int}$. This has clear implications for where further more accurate experimental data should be obtained.

6. An interval based approach

The approach as implemented above relies on developing statistical information about the data items (in the sampling and repeated solution of the stochastic model), often from few data points, and the expensive machinery of stochastic optimisation.

In the case of process development shown above data is often obtained as a measurement with error or uncertainty bounds. These bounds give an important indication of the uncertainty of the measurement but it is only with in depth knowledge can someone know whether the degree uncertainty in the measurement is going to have a significant effect in process development. The measurements are often taken by development chemists who are not involved in development of the manufacturing process and so while they will have a feeling for the effects on the chemistry they will not necessarily know the effect on the manufacturing process.

There is again a role for systematically incorporating the uncertainty into the development process using a model based approach. If the model based approach presented in figure 2 is used but with intervals rather than stochastic distributions a systematic approach can be developed. A

weakness is that if only intervals are used the approach could be very conservative but it should be able to indicate which uncertainty in which measurements have the most effect and whether the uncertainty in the design can be significantly improved by better measurement. If data is now provided as a measurement with error bounds (intervals), optimization could also be achieved by application of interval global optimisation algorithms.

Two important distinctions are identified in formulating flowsheeting problems. In the equation-oriented formulation the flowsheet is treated as a set of mass/energy balance equations that are solved simultaneously. The alternative sequential modular approach views the flowsheet as interconnected black boxes. Both approaches have their advantages; however the modular approach has a particular advantage in that it matches more closely the natural structure of the flowsheet. Modular approaches are in general more popular in the chemical industry. Using modular flowsheets built from general models Byrne and Bogle (2000) showed how interval methods could be used in conjunction with this type of system. Modular flowsheets are constructed with generic unit modules that can provide the interval bounds, linear bounds, derivatives and derivative bounds using extended arithmetic types. Using interval analysis and automatic differentiation as the arithmetic types, lower bounding information is used in a branch and bound network.

The approach shown in Figure 2 could be modified to exploit this interval information using interval optimization techniques to solve the optimization problems. Step 1 requires obtaining intervals instead of distributions of the model parameters. Step two defines instead a deterministic system but with intervals for the uncertain parameters (such as the activation energies) and uncertain outputs (such as yield in the example above). The sampling procedure is no longer necessary since the optimization is done in terms of the interval bounds only. Step 5 remains as for the stochastic problem and step six involves obtaining the globally optimal solution for the deterministic problem using the real data points. In step 4 the models are used to obtain the interval bounds on the output variables and a sensitivity analysis can be performed to determine the key predicted output uncertainties and hence reduce the dimensionality of the subsequent optimization problem. Finally an interval optimization problem should be solved to determine the optimal reduction in input uncertainty that will keep the output uncertainties within their desired limits.

This approach has the advantage of requiring only data and error bounds and can use the interval optimisation software that is available. Error bounds can be conservative and this approach will help to indicate when it would be most appropriate to really try and improve the accuracy of measurements by more careful procedures or by obtaining more sophisticated measuring equipment.

Conclusions

A systematic approach for incorporating uncertainty in process design has been presented. A stochastic optimisation problem is solved using distributions in the parameter uncertainties to determine where the key uncertainties in the data lie. This was applied to a multiphase batch reactor problem shown here and has also been applied to a pharmaceutical process involving 15 unit operations in sequence (Johnson). The methodology produced some clear recommendations about which measurements would best be improved to reduce the uncertainty in the output variables which are key for ensuring that the quality of the product is acceptable.

Since much data is often obtained from the laboratory with error bounds we have also discussed briefly how the problem could be cast as an interval optimisation problem which would determine where error bounds on particular data points were causing particular uncertainty in process development.

References

1. Basu P.K., R.A. Mack and J.M. Vinson, Consider a new approach to pharmaceutical process development, 1999, *Chem. Eng. Prog*., Vol. 95, No. 8, 82-90.
2. Byrne R.P. and Bogle I.D.L. (2000) Global optimisation of modular process flowsheets, Ind Eng Chem Res 2000 39 4296-4301
3. Diwekar U.M. and E.S. Rubin, Stochastic modelling of chemical processes, 1991, *Computers Chem. Engng,* Vol. 15, No. 2, 105-114.
4. Donaldson J.R. and R.B. Schnabel, Computational experience with confidence regions and confidence intervals for nonlinear least squares, 1987, *Technometrics*, Vol. 29, No. 1, 67-82.
5. Johnson (2003) Integrated Design Under Uncertainty for Pharmaceutical Processes. PhD thesis, University of London.
7. Hangos K. and I. Cameron, Process modelling and model analysis, 2001, Academic Press, London.
8. Sano T., T. Sugaya T and M. Kasai, Process improvement in the production of a pharmaceutical intermediate using a reaction calorimeter for studies of the reaction kinetics of amination of a bromopropyl compound, 1998, *Organic process research and development*, Vol. 2, 169-174.

Appendix

The deterministic model for the multiphase batch reactor

$$T_{iso,\max} \geq 7.06\left(t_{add}\right)^2 - 43.50\left(t_{add}\right) + 352.67$$

$$\frac{dm_A}{dt} = -k_1 \frac{m_A m_B}{V} - k_2 \frac{m_A m_C}{V}$$

$$\frac{dm_B}{dt} = \upsilon_{feed} \frac{m_{B0,feed}}{V_{B0,feed}} - k_1 \frac{m_A m_B}{V}$$

$$\frac{dm_C}{dt} = k_1 \frac{m_A m_B}{V} - k_2 \frac{m_A m_C}{V}$$

$$\frac{dm_D}{dt} = k_2 \frac{m_A m_C}{V}$$

$$k_{r,\text{int}} = A_{r,\text{int}} \exp\left(-\frac{Ea_{r,\text{int}}}{T_{iso}}\right) \qquad \text{for} \qquad r = 1,2$$

$$k_{1,diss} = A_{1,diss} \exp\left(-\frac{Ea_{1,diss}}{T_{iso}}\right)$$

$$k_{2,diss} = k_{2,\text{int}} \frac{k_{1,diss}}{k_{1,\text{int}}}$$

$$\beta_{diss} = \begin{cases} 1 & if \quad X_A \leq X_{diss} \\ 0 & if \quad X_A > X_{diss} \end{cases}$$

$$k_r = \beta_{diss} k_{r,diss} + \left(1 - \beta_{diss}\right) k_{r,\text{int}} \qquad \text{for} \qquad r = 1,2$$

$$\frac{dV}{dt} = \upsilon_{feed}\Big|_{t_0, t_{add}} = 0 \Big|_{t_{add}, t_f}$$

$$\upsilon_{feed} = \frac{V_{B0,feed}}{t_{add}}$$

$$T_{iso} \geq C_1\left(t_{add}\right)^2 - C_2\left(t_{add}\right) + C_3 \qquad \text{(safety constraint)}$$

$$X_A = \frac{m_{A0} - m_A}{m_{A0}}$$

$$\frac{dX_A}{dt} = \left(\frac{k_1 m_A m_B}{V m_{A0}} + \frac{k_2 m_A m_C}{V m_{A0}} \right) \leq 0.001 \quad hr^{-1}$$

$$T_{iso,\max} = C_1\left(t_{add}\right)^2 + C_2 t_{add} + C_3$$

Initial conditions (inside reactor)

$$m_{A0} = 1.075 \quad moles$$

$$m_{B0} = 0, \qquad m_{C0} = 0, \qquad m_{D0} = 0$$

$$V = 0.7 \quad dm^3$$

$$X_{A0} = 0$$

The subscripts diss and int denote dissolution and intrinsic kinetic controlled periods, and $k_{2,diss}$ is assumed to follow a similar temperature relationship as $k_{1,diss}$ relative to its intrinsic value.