Preprint Cambridge Centre for Computational Chemical Engineering ISSN 1473 – 4273

Influence of experimental observations on n-propylbenzene kinetic parameter estimates

Sebastian Mosbach and Markus Kraft [†]

released: 18 December 2013

[†] Department of Chemical Engineering
and Biotechnology
University of Cambridge
New Museums Site
Pembroke Street
Cambridge CB2 3RA
UK
Email: mk306@cam.ac.uk

Preprint No. 138

Keywords: n-Propylbenzene, regression influence diagnostics

Edited by
Computational Modelling Group
Department of Chemical Engineering and Biotechnology
University of Cambridge
New Museums Site
Pembroke Street
Cambridge CB2 3RA
United Kingdom

Fax: + 44 (0)1223 334796
E-Mail: c4e@cam.ac.uk
World Wide Web: `http://como.cheng.cam.ac.uk/`

Abstract

We calculate the derivatives of best estimates of kinetic parameters of an n-propylbenzene shock tube oxidation model, as determined by a weighted least-squares optimisation, with respect to experimental observations and compare these derivatives to some influence diagnostics based on omission of data points which are widely used in linear regression analysis. The considered data set comprises of 2378 measured concentrations of 37 stable species at various temperatures, pressures, and equivalence ratios. The methods studied are computationally affordable, as they require only a single optimisation and do not require the use of surrogates. We find that the diagnostics offer many insights into how individual observations influence parameter estimates, such as which observations determine which parameters to what extent. Additionally, the significance of non-linearities is investigated. While we observe that they can be of substantial importance to the derivatives, and improve the numerical conditioning of the involved matrix inversion, we find that results obtained from the linear omission-based diagnostics frequently agree, at least qualitatively, with those obtained from the derivatives.

Contents

1 Introduction

The goal of all experimentation is to gain knowledge about a physical system. Models can be regarded as a formalisation of such knowledge [20]. Assuming the form of a model, or several models, is fixed, the knowledge is then condensed in the parameters of the model(s). The following questions then arise naturally: Which parameters are determined by existing observations, and by which ones to what extent? And closely related to that, under which conditions should the next experiment be carried out in order to increase our knowledge most? The latter question is traditionally the subject of experimental design [15]. It is the former question we are concerned with in this paper, applied in particular to gas-phase combustion kinetic modelling.

One approach is uncertainty propagation [28, 32], which studies how errors in experimental measurements propagate through to model parameters (and responses). Those parameters with larger resulting errors are deemed less well determined by the data, and some methods allow inspection of the relative contribution of each data point, and its uncertainty, to the error in particular parameters. For example, in the Data Collaboration framework, a rigorous measure for data set consistency [16] has been developed. It has been used to analyse pairwise consistency of data set units, and for outlier identification.

There exists extensive literature on linear regression diagnostics. Several quantitative measures of the influence of observations on parameter estimates in linear regression, in order to detect influential data points, high-leverage points, and statistical outliers, have been proposed [12, 29]. An overview and comparison can be found in [3, 4]. One of the most popular ones is Cook's distance [6, 7, 9]. The basic idea underlying many of these measures is to carry out a regression using the full data set, and then repeating the regression with individual observations removed, for each of the observations. It is clear that this procedure becomes computationally prohibitive for large numbers of data points. For higher order exclusions, *i.e.* deletion of two or more observations, in order to study compensation effects for instance, the problem is aggravated by combinatorial explosion. In order to alleviate this, grouping of data points into sets is possible, but ultimately arbitrary and the selection of groupings may not be obvious. For linear models, however, it is possible to derive simplified expressions which require only a single regression to calculate the influence of removing each of the observations. The success and popularity of these diagnostics is largely due to this highly desirable property.

Another approach originates from sensitivity analysis in non-linear programming. Results on the so-called perturbation of the optimum for the more general situation of constrained optimisation can be traced back to at least 1968 [18, p. 34]. This approach, based on the implicit function theorem, is discussed extensively in [17]. The main results are expressions for the sensitivities of parameter estimates with respect to any other quantity in the objective function (or in constraints), which includes in particular experimental data. Essentially a special case of these methods was developed independently for unconstrained least-squares optimisation [14, 22], which was also applied in chemical kinetics [27, 34, 35]. Fiacco's work [18] appears to have gone unnoticed in the linear regression influence diagnostics community, even though methods based on sensitivities or derivatives are considered [4]. Neither the regression influence literature nor Fiacco's perturbation of the optimum appears to have been noticed in the area of combustion. Naturally, very similar ideas have also found application in optimal experimental design [11].

The purpose of this paper is to apply and compare a number of diagnostic techniques for the influence of experimental observations on parameter estimates to an extensive set of n-propylbenzene shock tube data. In particular, we give a theoretical exposition that is both accessible and shows the relationship between the methods. We test the methods by determining the influence of the observations on estimates of some Arrhenius pre-exponential factors in the kinetic mechanism used to model the data.

2 Theoretical background

We begin by presenting sensitivity analysis with respect to experiments [14, 27] in a way which illustrates the connection to the perturbation of the optimum [18], and to some linear regression diagnostics.

2.1 Problem formulation

Consider a vector ϑ of P model parameters whose 'best' values are estimated through minimisation of some objective function $\Phi(\tau, \vartheta)$, where $\tau \in \mathbb{R}^N$ denotes the vector of experimental observations:

$$\hat{\vartheta} = \underset{\vartheta}{\operatorname{argmin}}\, \Phi(\tau, \vartheta) \in \mathbb{R}^P \tag{1}$$

At a (local) minimum, we necessarily have

$$\frac{\partial \Phi}{\partial \vartheta_j} = 0. \tag{2}$$

The question is then how the optimum, *i.e.* the best estimate of the parameters, depends on the observations τ.

2.2 The Implicit Function Theorem

The Implicit Function Theorem states that, given a function

$$F : \mathbb{R}^N \times \mathbb{R}^P \to \mathbb{R}^P; \quad (\tau, \vartheta) \mapsto F(\tau, \vartheta)$$

and a point $(\tau_0, \vartheta_0) \in \mathbb{R}^N \times \mathbb{R}^P$ with $F(\tau_0, \vartheta_0) = 0$ and

$$\det\left(\frac{\partial F}{\partial \vartheta}(\tau_0, \vartheta_0)\right) \neq 0, \tag{3}$$

then ϑ can be expressed locally as a function of τ. In (3), $\frac{\partial F}{\partial \vartheta} \in \mathbb{R}^{P \times P}$ denotes the matrix of partial derivatives of F, with components $\left(\frac{\partial F}{\partial \vartheta}\right)_{jk} = \frac{\partial F_j}{\partial \vartheta_k}$. The derivatives of ϑ

with respect to τ can then be obtained by differentiating $F(\tau, \vartheta) = 0$ with respect to τ whilst considering ϑ as a function of τ and applying the chain rule. The resulting expression can be solved for the sought-after derivatives of the parameters with respect to the observations:

$$\frac{\partial \vartheta_j}{\partial \tau_i} = -\sum_{k=1}^{P} \left(\frac{\partial F}{\partial \vartheta}^{-1}\right)_{jk} \frac{\partial F_k}{\partial \tau_i} \tag{4}$$

2.3 Application

Now, choosing

$$F_j(\tau, \vartheta) = \frac{\partial \Phi}{\partial \vartheta_j}(\tau, \vartheta) \tag{5}$$

allows calculating the derivatives of the model parameters ϑ with respect to experimental observations τ via (4).

We note that any constant added to F does not affect the derivatives (4), implying that any non-extremal point on the objective function surface can be used as well. It is conceivable, though, that a non-extremal point has a different dependence on (and hence different derivatives) the experimental observations than the (nearest) minimum. We furthermore note that for the choice (5) condition (3) becomes $\det H \neq 0$ with $H_{jk} = \frac{\partial^2 \Phi}{\partial \vartheta_j \partial \vartheta_k}$, *i.e.* the determinant of the Hessian must not vanish. If the considered point is at the bottom or on the slope of a valley [19] one might expect the Hessian to be (at least close to) degenerate, so the procedure outlined may be numerically ill-conditioned, but this turns out to be unproblematic.

Consider now the objective function

$$\Phi(\tau, \vartheta) = \big(\tau - \eta(\vartheta)\big)^{\top} \Sigma^{-1} \big(\tau - \eta(\vartheta)\big), \tag{6}$$

with $\Sigma \in \mathbb{R}^{N \times N}$ denoting the (symmetric and positive definite) covariance matrix of the responses $\eta \in \mathbb{R}^N$. We note that, for an objective function, it does not matter whether a model is single or multi-response – the difference is merely notational. Then, (5) implies

$$F_j(\tau, \vartheta) = -2 \sum_{li} \frac{\partial \eta_l}{\partial \vartheta_j} (\Sigma^{-1})_{li} \big(\tau_i - \eta_i(\vartheta)\big) \tag{7}$$

and

$$\frac{\partial F_j}{\partial \tau_i} = -2 \sum_{l} \frac{\partial \eta_l}{\partial \vartheta_j} (\Sigma^{-1})_{li} \tag{8}$$

and

$$\frac{\partial F_j}{\partial \vartheta_k} = 2 \sum_{li} \frac{\partial \eta_l}{\partial \vartheta_j} (\Sigma^{-1})_{li} \frac{\partial \eta_i}{\partial \vartheta_k} - 2 \sum_{li} \frac{\partial^2 \eta_l}{\partial \vartheta_j \partial \vartheta_k} (\Sigma^{-1})_{li} \big(\tau_i - \eta_i(\vartheta)\big) \tag{9}$$

One may be concerned that (4) appears to be local with respect to τ, and thereby be led to consider global methods [32]. However, it should be noted that (7) is an affine-linear function in τ. Further, (9) is affine-linear and (8) is constant with respect to τ, the latter implying $\frac{\partial^2 F}{\partial \tau_i \partial \tau_l} = 0$.

We notice that $\frac{\partial \eta}{\partial \vartheta}^\top \Sigma^{-1} \frac{\partial \eta}{\partial \vartheta}$ is the (Fisher) information matrix familiar from experimental design [15, 24].

The second term in (9) is zero if the residuals are zero, *i.e.* if the model agrees with experiment perfectly, even for a non-linear model. As we shall see below, this second term, originating from model non-linearities due to the presence of the second order derivatives, can however also be dominant.

2.4 Linear regression influence diagnostics

In order to introduce the diagnostics and to see the relationship to (4), it is instructive to recall the general linear model, *i.e* a model which is linear in its model parameters ϑ but not necessarily in the process conditions ξ. That means its (single) response is given by $g^\top(\xi)\vartheta$, with g a vector of arbitrary functions of the process conditions, *i.e.* $g(\xi) = \big(g_1(\xi), g_2(\xi), \dots, g_P(\xi)\big)^\top$. It is assumed that the relationship between the collections of experimental and model responses is well-described by $\tau = X\vartheta + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \Sigma)$. The design matrix $X \in \mathbb{R}^{N \times P}$ has components $X_{ij} = g_j(\xi^{(i)})$, where $\xi^{(i)}$ is the process condition at which the i^{th} observation has been made. Then, the maximum likelihood estimate $\hat{\vartheta}$ of ϑ can be obtained from (1) and (6) as

$$\hat{\vartheta} = \big(X^\top \Sigma^{-1} X\big)^{-1} X^\top \Sigma^{-1} \tau. \tag{10}$$

Furthermore, ϑ has covariance matrix $\big(X^\top \Sigma^{-1} X\big)^{-1}$. From (10), we obtain the derivatives of the parameter estimates with respect to experimental observations as

$$\frac{\partial \hat{\vartheta}}{\partial \tau} = \big(X^\top \Sigma^{-1} X\big)^{-1} X^\top \Sigma^{-1}. \tag{11}$$

Now, realising that for the linear model $\frac{\partial \eta}{\partial \vartheta} = X$ and that the second derivatives of η with respect to ϑ vanish, we obtain for (8) $\frac{\partial F}{\partial \tau} = -2X^\top \Sigma^{-1}$, and for (9) $\frac{\partial F}{\partial \vartheta} = 2X^\top \Sigma^{-1} X$. Thus, for the linear model (4) is identical to (11).

We now consider the effect of deleting a single observation from a data set. One of the simplest statistics for assessing this effect is called DFBETA_i [2, p. 13]. It is defined as the difference between the best estimate of the parameters and the best estimate with the i^{th} data point removed:

$$\mathsf{DFBETA}_i := \hat{\vartheta} - \hat{\vartheta}_{-i} \in \mathbb{R}^P \tag{12}$$

Definition (12) is meaningful for any non-linear model, however, direct use would require one regression or optimisation per data point, which is in general infeasible. Simplified formulae which do not require multiple additional regressions can be derived as follows, but these are strictly valid only for linear models. For this, it is necessary to restrict from a Generalised Least Squares (GLS) problem, *i.e.* with Σ an arbitrary positive definite matrix, to Weighted Least Squares (WLS), which assumes $\Sigma_{il} = \sigma_i^2 \delta_{il}$, *i.e.* uncorrelated but possibly heteroskedastic data. While it is possible to transform a GLS into a WLS problem, omitting the i^{th} data point from the GLS is not equivalent to omitting the i^{th} data point from the transformed problem. Defining $\mathcal{I} := X^\top \Sigma^{-1} X$, we can write

$$\hat{\vartheta} = \mathcal{I}^{-1} X^\top \Sigma^{-1} \tau \qquad \text{and} \qquad \hat{\vartheta}_{-i} = \mathcal{I}_{-i}^{-1} X_{-i}^\top \Sigma_{-i}^{-1} \tau_{-i}, \tag{13}$$

where $X_{-i} \in \mathbb{R}^{(N-1)\times P}$ denotes X with its i^{th} row removed, $\Sigma_{-i} \in \mathbb{R}^{(N-1)\times(N-1)}$ denotes Σ with its i^{th} row and column removed, and $\mathcal{I}_{-i} := X_{-i}^\top \Sigma_{-i}^{-1} X_{-i}$. We then obtain

$$\mathcal{I}_{-i} = \mathcal{I} - \sigma_i^{-2} x^{(i)} x^{(i)\top},$$

where $x^{(i)} \in \mathbb{R}^P$ is the *column* vector containing the i^{th} *row* of X. The key step is then to express $\mathcal{I}_{-i}^{-1}$ in terms of $\mathcal{I}^{-1}$ by means of the Sherman-Morrison-Woodbury formula. Inserting the result into (12) then, after some algebra, yields

$$\hat{\vartheta} - \hat{\vartheta}_{-i} = \frac{\mathcal{I}^{-1} x^{(i)}}{\sigma_i^2 - x^{(i)\top} \mathcal{I}^{-1} x^{(i)}} \big(\tau_i - x^{(i)\top} \vartheta\big). \tag{14}$$

This formula allows calculating DFBETA_i for all observations i with only a single regression involved. We note that (14) is proportional to the residuals, so will be zero whenever the model agrees with experiment.

A scaled variant $\mathsf{DFBETAS}_i$ can be defined by dividing each of the components of DFBETA_i by the square root of the variance of the corresponding parameter, or, perhaps more usefully, by the square root of the empirical variance with the i^{th} data point removed [2, p. 13]. The motivation for such scaling is that a change in a parameter, and the significance thereof, should be assessed with respect to the magnitude of its uncertainty. For this reason, in this work, we consider

$$B_{ji} := \frac{(\hat{\vartheta} - \hat{\vartheta}_{-i})_j}{\sqrt{(\mathcal{I}^{-1})_{jj}}} \tag{15}$$

as one of the diagnostics, with the numerator given by (14).

Another statistic can be obtained by differentiating the best parameter estimates (10) with respect to the weights in the objective function [2, pp. 24 & 66], *i.e.* the diagonal entries of the (WLS) covariance matrix Σ (see also [26]). The resulting expression is essentially (11) multiplied by the residuals. Up to normalisation, this is the same as $\mathsf{DFBETA(S)}$, which is why it has been suggested as an alternative for the latter.

One of the most well-known influence diagnostics is Cook's distance. The motivation for its definition originates from joint confidence regions of the parameters. Joint $100(1-\alpha)\%$ confidence ellipsoids can be defined as

$$(\hat{\vartheta} - \vartheta)^\top X^\top \Sigma^{-1} X (\hat{\vartheta} - \vartheta) \leq P s^2 F(P, N-P, 1-\alpha),$$

with $s^2 := \frac{1}{N-P} r^\top \Sigma^{-1} r = \Phi/(N-P)$, residuals $r := \tau - X\hat{\vartheta}$, and $F(P, N-P, 1-\alpha)$ the $1-\alpha$ point of the F-distribution. See [13, pp. 94 & 108] and [19]. Cook's distance is defined as

$$D_i := \frac{(\hat{\vartheta} - \hat{\vartheta}_{-i})^\top X^\top \Sigma^{-1} X (\hat{\vartheta} - \hat{\vartheta}_{-i})}{P s^2}. \tag{16}$$

Large values of D_i mean large differences in the parameter estimates when case i is deleted. The original definition [6, 7] was given for ordinary least squares, and later generalised to weighted [9, p. 209]. While definition (16) is applicable to general least squares, simplified formulas require a WLS formulation. As with (12), direct use of definition (16) would require one optimisation per data point, which is infeasible in general.

Using (14), we can obtain a simplified expression for Cook's distance (16) which requires only a single regression for all observations:

$$D_i = \frac{v_{ii}}{Ps^2}\left(\frac{r_i}{\sigma_i^2 - v_{ii}}\right)^2, \tag{17}$$

where $v_{ii} := x^{(i)\top}\mathcal{I}^{-1}x^{(i)}$. We use Eqn. (17) as one of the diagnostics in this paper.

We note that Cook's distance is a measure of the *overall* influence of an observation and does not allow studying parameters individually. Influence measures for the deletion of data points on subsets of parameters, which includes individual ones as a special case, have been proposed (see [3] for an overview). A notable example is the following. Cook and Weisberg [8] have presented a version of Cook's distance which quantifies the effect of deletion of (sets of) data points on subsets of the estimated parameters. As the original distance, it is only applicable to linear models, but in addition requires one extra regression per considered parameter subset. That is, if all parameters are of interest individually, P additional regressions are required.

Measures quantifying non-linearity based on local curvature, distinguishing intrinsic from parametrisation-related curvature, have been proposed [1, 30], but it is beyond the scope of the paper to explore this.

3 A note on implementation

Calculating (9) numerically involves calculating the second derivatives of the model responses. This can be done analytically when polynomial surrogates are used for example, but more generally requires either the direct method of integrating the set of ordinary differential equations governing the second order sensitivity coefficients or finite-difference approximations [27, 34]. In this work, we employ the latter.

The standard method is to expand different perturbations of a function f into their multivariate Taylor series and then take appropriate linear combinations. This yields for example

$$\frac{\partial^2 f}{\partial x \partial y} \approx \frac{f_{++} + f_{--} - f_{+\cdot} - f_{-\cdot} - f_{\cdot+} - f_{\cdot-} + 2f}{2\Delta x \Delta y} \tag{18}$$

with a second-order error term, where we have used notations such as $f_{++} := f(x_0 + \Delta x, y_0 + \Delta y)$ and $f_{-\cdot} := f(x_0 - \Delta x, y_0)$. While stencil (18) may not be the simplest, it involves only *two* function evaluations with perturbations in both variables per pair. This implies that for this stencil a total of $1 + 2P + 2\binom{P}{2} = P^2 + P + 1$ evaluations are required. Finite difference scheme (18) also has the added advantage that the univariate second-order derivatives as well as the first-order derivatives with a higher order in the error term can be approximated without further evaluations. For the univariate second-order derivatives appearing in (9), we use the standard central difference scheme.

4 Experimental data

The data set considered here [21] (available online) was obtained using the high-pressure single-pulse shock tube at the University of Illinois at Chicago [31, 33]. Process condition variables comprise of initial temperature, initial pressure, initial composition, and reaction time. In all cases, the initial mixture is composed of n-propylbenzene, oxygen, and argon. The system is highly dilute, and hence can be considered isothermal. An overview of the conditions is given in Table 1. Concentrations of stable species were determined via gas chromatography in gases sampled from the shock tube. Measured species include O_2, CO, CO_2, 14 saturated as well as unsaturated aliphatic hydrocarbons, 16 aromatic hydrocarbons with up to three rings including benzene, toluene, phenylacetylene, 1-propenylbenzene, 2-propenylbenzene, and n-propylbenzene, and four oxygenated aromatics. Measurement errors range between 1.7% and 25%, with an average of about 12%. In total, there are 2378 experimental observations of 37 species at 74 points in process-condition space.

Table 1: *Overview of experimental process conditions.*

p_{avg} [atm]	T [K]	Fuel [ppm]	Φ	t_{reac} [ms]
28	907-1551	86	0.54	1.40-2.05
51	959-1558	90	0.55	1.27-1.90
49	838-1635	90	1.0	1.21-1.95
24	905-1669	89	1.9	1.36-2.93
52	847-1640	90	1.9	1.26-1.95

5 Model and optimisation

The shock tube is modelled as a homogeneous, adiabatic, constant pressure reactor. As software to solve the governing equations, which we shall not repeat here, *k*inetics v8.0 [5] was employed. A chemical kinetic mechanism containing 191 species and 1127 reactions [10, 21, 23] was used. All experimentally measured species mentioned above are present in the mechanism.

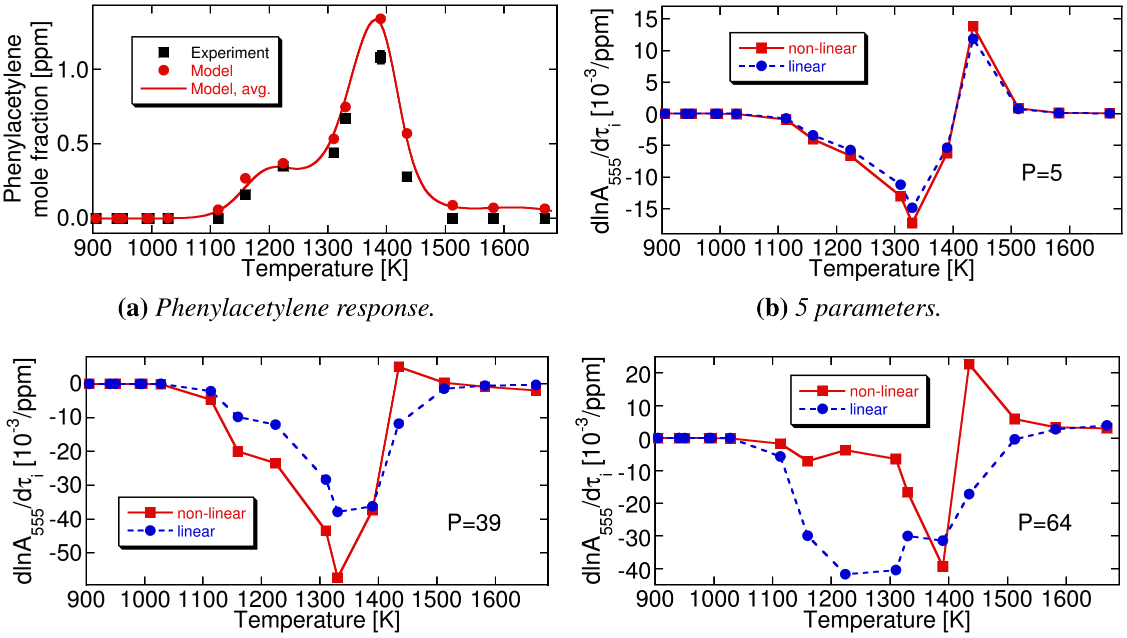
In a previous paper [25], we optimised 64 Arrhenius pre-exponential factors, which were chosen through conventional sensitivity analysis, against the above data set.

6 Results

We conduct the influence analysis on three parameter sets: one including all of the $P = 64$ parameters used for the optimisation, one with a subset of $P = 39$ parameters, and one with an even smaller subset of $P = 5$ parameters. The corresponding simulations involve P^2+P+1 objective function evaluations, *i.e* 4161, 1561, and 31 respectively. We find that

Table 2: *Contribution of non-linearities to derivatives of best estimates with respect to observations (Eqn. 4).*

	$P = 5$	$P = 5$	$P = 39$	$P = 64$
Quantity	$\vartheta = A$	$\vartheta = \ln A$	$\vartheta = \ln A$	$\vartheta = \ln A$
$\Vert\frac{\partial F}{\partial \vartheta} - 2\mathcal{I}\Vert / \Vert\frac{\partial F}{\partial \vartheta}\Vert$	4.9%	3.8%	10.5%	11.1%
cond $\frac{\partial F}{\partial \vartheta}$	1.1×10^{32}	179	1.5×10^{5}	1.3×10^{5}
cond $\mathcal{I}$	1.0×10^{32}	175	7.6×10^{7}	1.4×10^{8}



(a) *Phenylacetylene response.*

(b) *5 parameters.*

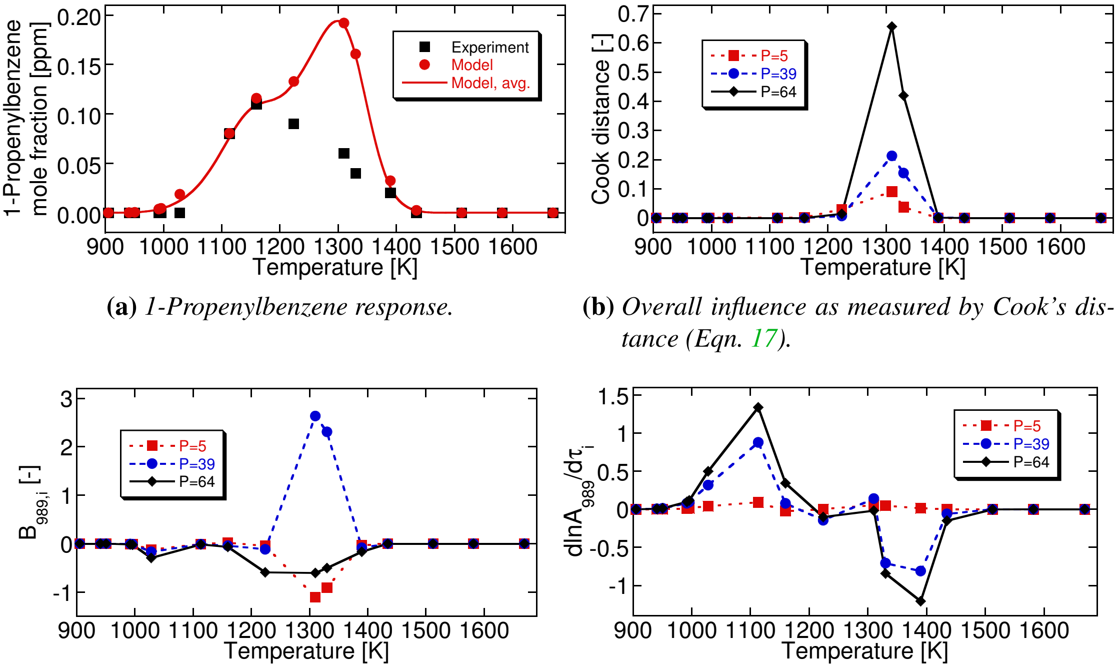
(c) *39 parameters.*

(d) *64 parameters.*

Figure 1: *Influence on pre-exponential factor of reaction 555 as measured by derivatives of best estimates with respect to observations (Eqn. 4) for phenylacetylene data at an average pressure of* 24 *atm and* $\Phi = 1.9$. *The full expressions (labelled 'non-linear') as well as the ones with the second derivatives omitted (labelled 'linear') are shown for the three considered parameter sets.*

both the information matrix and the Hessian matrix are typically highly ill-conditioned. In the 5-parameter case, their condition numbers, *i.e.* the ratio of the largest to the smallest eigenvalue, denoted by $\mathrm{cond}(\cdot)$, are of order 10^{32} – runs with more parameters failed altogether. This is dramatically improved by considering the logarithms of the model parameters instead, *i.e.* $\vartheta_j = \ln A_j$. Details are given in Table 2.

In order to assess the significance of non-linearities, we compare the two terms in (9) by means of the Frobenius norm, *i.e.* the square root of the sum of the squares of all matrix elements, denoted by $\Vert \cdot \Vert$, and matrix condition numbers. We find that the second term, which involves the second derivatives, *i.e.* the contribution from model non-linearities, is small (up to $\approx 11\%$ Frobenius norm), but it can significantly improve the conditioning of



(a) *1-Propenylbenzene response.*

(b) *Overall influence as measured by Cook's distance (Eqn. 17).*

(c) *Influence on pre-exponential factor of reaction 989 as measured by B_{ji} (Eqn. 15).*

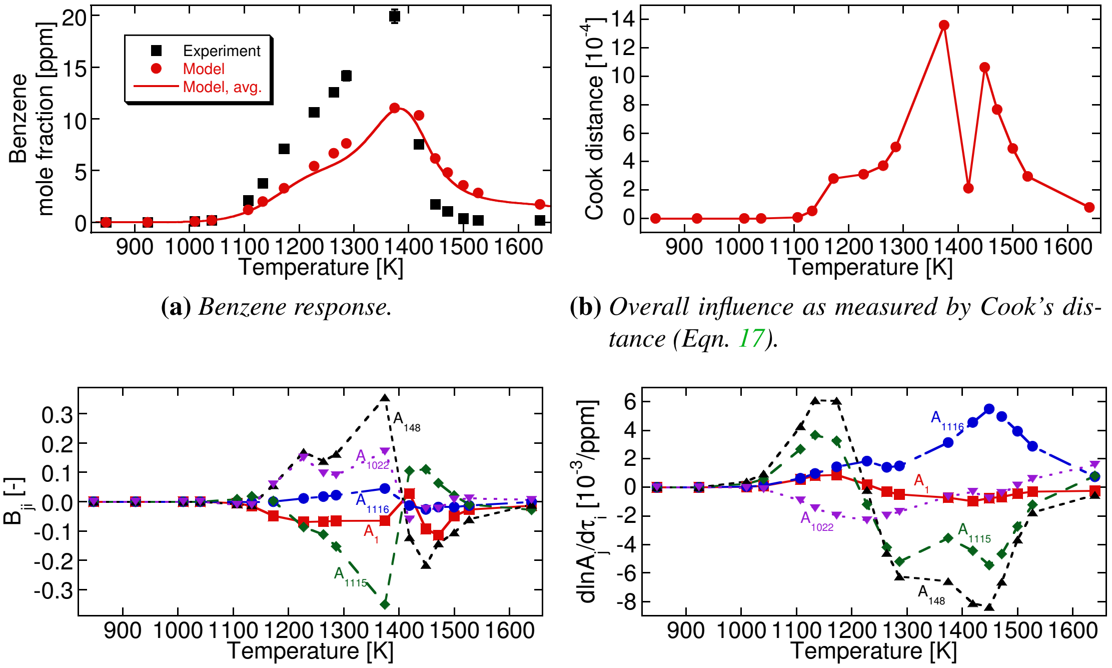
(d) *Influence on pre-exponential factor of reaction 989 as measured by derivatives of best estimates with respect to observations (Eqn. 4).*

Figure 2: *Variation of influence measures with considered parameter set for 1-propenylbenzene observations at an average pressure of* 24 *atm and* $\Phi = 1.9$.

the Hessian matrix (by up to three orders of magnitude, see Table 2).

Figure 1 shows the derivatives of the best estimates of pre-exponential of reaction 555 with respect to phenylacetylene observations for each of the three considered parameter sets. Two versions are shown for each set: The full expression, *i.e.* Eqn. (4) with (9), labelled 'non-linear', and the linearised one, *i.e.* with the second term in (9) involving the second derivatives omitted, labelled 'linear'. While there is virtually no difference between the two for the 5-parameter set, differences become appreciable for $P = 39$ and considerable for $P = 64$.

Figure 2 shows the three considered influence measures, Eqns. (17), (15), and (4), for each of the three parameter sets. While for each of the diagnostics there is some variation from set to set, it is mostly in magnitude – the relative significance of observations is largely similar. For Cook's distance as well as B_{ji} (Eqn. 15), the proportionality to the residuals is clearly visible, whereas the derivatives (Fig. 2d) indicate that observations at about 1100 K are equally important. We note that some dependence on the parameter set should be expected since even for a linear model, the derivatives of a particular parameter are affected by the presence of other parameters. Similarly, the Hessian matrix, *i.e.* the curvature of the objective function, which appears in (4), is of course dependent on the choice of parameters. One should also be aware that, more generally, the results depend on the local geometry of the objective function surface and hence on the choice of objective function and the chosen local minimum.



(a) *Benzene response.*

(b) *Overall influence as measured by Cook's distance (Eqn. 17).*

(c) *Influence as measured by B_{ji} (Eqn. 15).*

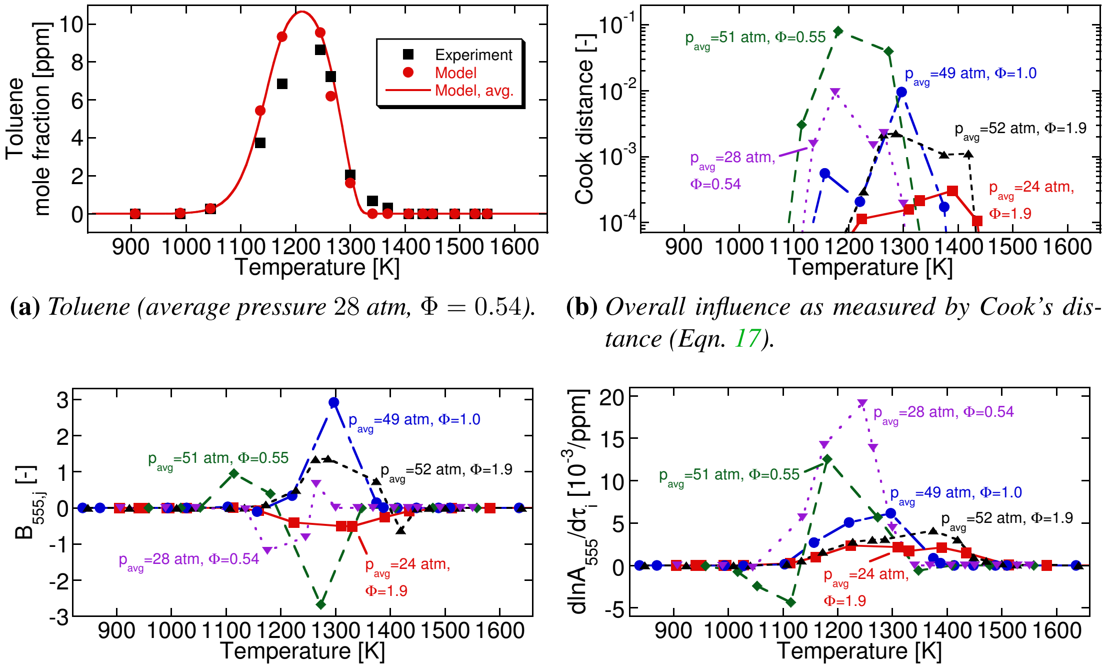
(d) *Influence as measured by derivatives of best estimates with respect to observations (Eqn. 4).*

Figure 3: *Influence on pre-exponential factors of five selected reactions for benzene observations at an average pressure of* 52 atm *and* $\Phi = 1.9$ *(*$P = 64$*).*

The largest contribution to the value of the objective function for the original model prior to optimisation comes from 1- and 2-propenylbenzene due to a relatively large difference between experiment and model, and relatively small experimental error bars (implying a large weight). The most influential observation overall, interestingly, by Cook distance is also 2-propenylbenzene, followed by phenylacetylene and 1-propenylbenzene. According to the derivatives, for most parameters, it is 1-propenylbenzene, and for the remaining ones 2-propenylbenzene (all at intermediate temperatures).

Figure 3 shows the relative influence of benzene observations at an average pressure of 52 atm and $\Phi = 1.9$ on the pre-exponential factors of the five most sensitive reactions. In contrast to the derivatives, the omission-based diagnostics do not deem the data points around 1150 K to be significant. Once again, as in Fig. 2, this is a reflection of the fact that Cook's distance and B_{ji} are mostly proportional to the residuals.

Figure 4 shows the influence of all observations of toluene overall and on the pre-exponential factor of reaction 555, grouped by pressure and equivalence ratio. All three diagnostics agree that the observations at lean mixtures are most influential, and that measurements at higher pressures are more influential (for fixed Φ).



(a) *Toluene (average pressure 28 atm, $\Phi = 0.54$).*

(b) *Overall influence as measured by Cook's distance (Eqn. 17).*

(c) *Influence on pre-exponential factor of reaction 555 as measured by B_{ji} (Eqn. 15).*

(d) *Influence on pre-exponential factor of reaction 555 as measured by derivatives of best estimates with respect to observations (Eqn. 4).*

Figure 4: *The three considered influence measures for all observations of toluene ($P = 64$).*

7 Conclusions

We have applied and compared some diagnostic measures for the influence of experimental observations on parameter estimates to an n-propylbenzene shock tube data set. For the model and data considered here, we found that non-linearities can be significant, and in particular improve the conditioning of the matrix inversion involved in calculating the derivatives with respect to observations. We furthermore found that, in spite of that, the three considered diagnostics give roughly similar results. Given that omission of an observation is much more 'drastic' than perturbation of the value of an observation, which is what derivatives measure, the similarity of the diagnostics is perhaps surprising. More generally, the additional information available through the derivatives needs to be weighed against the computational cost associated with their evaluation, which may be substantial for large numbers of parameters. The two linear diagnostics considered in this paper can be calculated from the first-order sensitivities of responses with respect to model parameters alone. As none of the diagnostics stands out as clearly superior to the others, we conclude overall that it is advisable to consider multiple ones. A natural next step is to compare the diagnostics with model-based experimental design techniques such as [24].

Acknowledgement

This publication is made possible by the Singapore National Research Foundation under its Campus for Research Excellence And Technological Enterprise (CREATE) programme.

References

[1] D. M. Bates and D. G. Watts. Relative curvature measures of nonlinearity. *Journal of the Royal Statistical Society B*, 42(1):1–25, 1980.

[2] D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression Diagnostics : Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, 1980.

[3] S. Chatterjee and A. S. Hadi. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–393, 1986. doi:10.1214/ss/1177013622.

[4] S. Chatterjee and A. S. Hadi. *Sensitivity Analysis in Linear Regression*. John Wiley & Sons, New York, 1988.

[5] cmcl innovations. *k*inetics: the chemical kinetics model builder, version 8.0, 2013. http://www.cmclinnovations.com/kinetics/.

[6] R. D. Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977. Stable URL: http://www.jstor.org/stable/1268249.

[7] R. D. Cook. Influential observations in linear regression. *Journal of the American Statistical Association*, 74(365):169–174, 1979. Stable URL: http://www.jstor.org/stable/2286747.

[8] R. D. Cook and S. Weisberg. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508, 1980. doi:10.1080/00401706.1980.10486199. Stable URL: http://www.jstor.org/stable/1268187.

[9] R. D. Cook and S. Weisberg. *Residuals and Influence in Regression*. Chapman and Hall, New York, 1982.

[10] P. Dagaut, A. Ristori, A. El Bakali, and M. Cathonnet. Experimental and kinetic modeling study of the oxidation of n-propylbenzene. *Fuel*, 81(2):173–184, 2002. doi:10.1016/S0016-2361(01)00139-9.

[11] H. Dette, V. B. Melas, and P. Shpilev. Optimal designs for estimating the derivative in nonlinear regression. *Statistica Sinica*, 21:1557–1570, 2011. doi:10.5705/ss.2009.202.

[12] N. R. Draper and J. A. John. Influential observations and outliers in regression. *Technometrics*, 23(1):21–26, 1981. doi:10.1080/00401706.1981.10486232. Stable URL: http://www.jstor.org/stable/1267971.

[13] N. R. Draper and H. Smith. *Applied Regression Analysis*. John Wiley & Sons, New York, 2nd edition, 1981.

[14] L. Eno, J. G. B. Beumee, and H. Rabitz. Sensitivity analysis of experimental data. *Applied Mathematics and Computation*, 16(2):153–163, 1985. doi:10.1016/0096-3003(85)90005-0.

[15] V. V. Fedorov. *Theory of Optimal Experiments*. Academic Press, New York, 1972.

[16] R. Feeley, P. Seiler, A. Packard, and M. Frenklach. Consistency of a reaction dataset. *Journal of Physical Chemistry A*, 108(44):9573–9583, 2004. doi:10.1021/jp047524w.

[17] A. V. Fiacco. *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, volume 165 of *Mathematics in Science and Engineering*. Academic Press, New York, 1983.

[18] A. V. Fiacco and G. P. McCormick. *Nonlinear Programming – Sequential Unconstrained Minimization Techniques*. Classics in Applied Mathematics. SIAM, 1990.

[19] M. Frenklach. Modeling. In W. C. Gardiner, editor, *Combustion Chemistry*, chapter 7, pages 423–453. Springer Verlag, New York, 1984.

[20] M. Frenklach. Transforming data into knowledge – Process Informatics for combustion chemistry. *Proceedings of the Combustion Institute*, 31(1):125–140, 2007. doi:10.1016/j.proci.2006.08.121.

[21] S. Gudiyella and K. Brezinsky. High pressure study of n-propylbenzene oxidation. *Combustion and Flame*, 159(3):940–958, 2012. doi:10.1016/j.combustflame.2011.09.013.

[22] R. E. Kalaba and K. Spingarn. Sensitivity of parameter estimates to observations, system identification, and optimal inputs. *Applied Mathematics and Computation*, 7 (3):225–235, 1980. doi:10.1016/0096-3003(80)90045-4.

[23] T. A. Litzinger, K. Brezinsky, and I. Glassman. Reactions of n-propylbenzene during gas phase oxidation. *Combustion Science and Technology*, 50(1-3):117–133, 1986. doi:10.1080/00102208608923928.

[24] S. Mosbach, A. Braumann, P. L. W. Man, C. A. Kastner, G. P. E. Brownbridge, and M. Kraft. Iterative improvement of Bayesian parameter estimates for an engine model by means of experimental design. *Combustion and Flame*, 159(3):1303–1313, 2012. doi:10.1016/j.combustflame.2011.10.019.

[25] S. Mosbach, J. H. Hong, G. P. E. Brownbridge, M. Kraft, S. Gudiyella, and K. Brezinsky. Bayesian error propagation for a kinetic model of n-propylbenzene oxidation in a shock tube. 2013. Submitted for publication.

[26] W. Polasek. Regression diagnostics for general linear regression models. *Journal of the American Statistical Association*, 79(386):336–340, 1984. Stable URL: http://www.jstor.org/stable/2288273.

[27] H. Rabitz, M. Kramer, and D. Dacol. Sensitivity analysis in chemical kinetics. *Annual Review of Physical Chemistry*, 34:419–461, 1983. doi:10.1146/annurev.pc.34.100183.002223.

[28] M. Sander, R. I. A. Patterson, A. Braumann, A. Raj, and M. Kraft. Developing the PAH-PP soot particle model using process informatics and uncertainty propagation. *Proceedings of the Combustion Institute*, 33(1):675–683, 2011. doi:10.1016/j.proci.2010.06.156.

[29] R. Schall and T. T. Dunne. Influential variables in linear regression. *Technometrics*, 32(3):323–330, 1990. doi:10.1080/00401706.1990.10484685. Stable URL: http://www.jstor.org/stable/1269109.

[30] G. A. F. Seber and C. J. Wild. *Nonlinear Regression*. John Wiley & Sons, 2003.

[31] W. Tang and K. Brezinsky. Chemical kinetic simulations behind reflected shock waves. *International Journal of Chemical Kinetics*, 38(2):75–97, 2006. doi:10.1002/kin.20134.

[32] A. S. Tomlin. The role of sensitivity and uncertainty analysis in combustion modelling. *Proceedings of the Combustion Institute*, 34(1):159–176, 2013. doi:10.1016/j.proci.2012.07.043.

[33] R. S. Tranter, K. Brezinsky, and D. Fulle. Design of a high-pressure single pulse shock tube for chemical kinetic investigations. *Review of Scientific Instruments*, 72 (7):3046–3054, 2001. doi:10.1063/1.1379963.

[34] T. Turányi. Sensitivity analysis of complex kinetic systems. Tools and applications. *Journal of Mathematical Chemistry*, 5(3):203–248, 1990. doi:10.1007/BF01166355.

[35] R. A. Yetter, H. Rabitz, F. L. Dryer, R. G. Maki, and R. B. Klemm. Evaluation of the rate constant for the reaction OH+H_2CO: Application of modeling and sensitivity analysis techniques for determination of the product branching ratio. *Journal of Chemical Physics*, 91(7):4088–4097, 1989. doi:10.1063/1.456838.