Estimation, model discrimination, and experimental design for implicitly given nonlinear models of enzyme catalyzed chemical reactions

Anna Siudak
Carl von Ossietzky University Oldenburg

Eric von Lieres
Research Centre Jülich

Christine H. Müller*
University Kassel

March 28, 2008

Abstract

Many nonlinear models as e.g. models of chemical reactions are described by systems of differential equations which have no explicit solution. In such cases the statistical analysis is much more complicated than for nonlinear models with explicitly given response functions. Numerical approaches need to be applied in place of explicit solutions. This paper describes how the analysis can be done when the response function is only implicitly given by differential equations. It is shown how the unknown parameters can be estimated and how these estimations can be applied for model discrimination and for deriving optimal designs for future research. The methods are demonstrated with a chemical reaction catalyzed by the enzyme Benzaldehyde lyase.

1 Introduction

Many scientific processes can be approximately described by mathematical models. If one has to deal with material fluxes or other continuously time-dependent changes or reactions, these models often contain implicitly given nonlinear differential equations. The advantage of expressing for example biochemical or physical systems by mathematical models is, that various statistical tools for estimating unknown model-parameters or discriminating between alternative models are available. Some of these statistical methods shall be discussed in this paper concerning their applicability for implicitly given nonlinear models. They are applied to an example of recent biocatalytic research, where the model is a differential equation model describing enzyme catalyzed reactions.

*Research supported by the SFB/TR TRR 30 Project D6

The example is given by an in-vitro catalysis, performed by the enzyme Benzaldehyde lyase. The Benzaldehyde lyase (BAL) is an enzyme which catalyzes among other reactions the production of benzoins from aromatic aldehydes and the compounding of aromatic aldehydes with aliphatic aldehydes. The modes of operation of the BAL in the catalysis are not completely understood up to now.

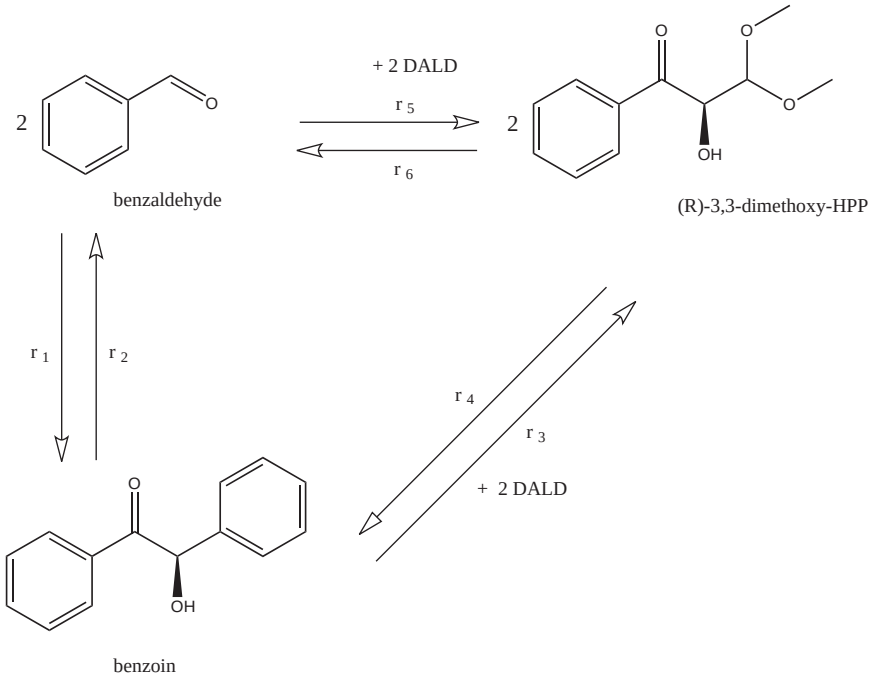

Figure 1: First model for the reaction from BA to DHPP [Reaction Model A]

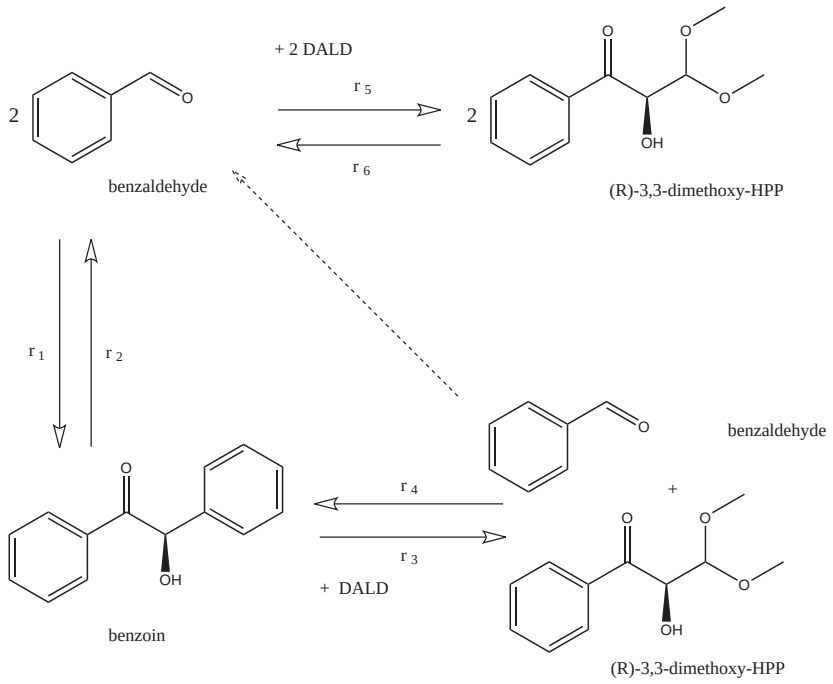

Figure 2: Second model for the reaction from BA to DHPP [Reaction Model B]

In this paper, we consider the enzymatical transformation of benzaldehyde (BA) and dimethoxyaldehyde (DALD) to the end product (R)-3,3-dimethoxy-1-phenyl-2-

hydroxypropan-1-one (DHPP). Considering the function of the BAL in other reactions, there are two most preferable reaction models for the transformation given by Figure 1 and Figure 2 (see [DEH+02] and [Küh07]).

The two models differ in particular in the reaction of the intermediate product benzoin (BZ) to DHPP: two additional molecules DALD are required in Reaction Model A while only one molecule DALD is needed in Reaction Model B. In addition the produced benzaldehyde in Reaction Model B can be converted anew (hinted at in Fig. 2 by a dashed arrow). This leads to the following reaction equations, which reduce the reaction system to its most necessary elements:

Reaction Model A

$$\begin{aligned} \frac{d\, c_{BA}}{dt} &= -2r_1, && (1)\\ \frac{d\, c_{BZ}}{dt} &= r_1 - r_3, && (2)\\ \frac{d\, c_{DHPP}}{dt} &= 2r_3, && (3) \end{aligned}$$

Reaction Model B

$$\begin{aligned} \frac{d\, c_{BA}}{dt} &= -2r_1 + r_3, && (4)\\ \frac{d\, c_{BZ}}{dt} &= r_1 - r_3, && (5)\\ \frac{d\, c_{DHPP}}{dt} &= r_3, && (6) \end{aligned}$$

with

$$\begin{aligned} r_1 &= V_{max,BABZ} * \left(\frac{c_{BA}}{K_{m,BABZ} + c_{BA}} \right)^2, && (7)\\ r_3 &= V_{max,BZDHPP} * \frac{c_{BZ}}{K_{m,BZDHPP} + c_{BZ}}. && (8) \end{aligned}$$

The main parts of the models are given by Michaelis-Menten kinetics. The squared term in (7) is caused by the fact that two molecules BA are transformed to one molecule BZ. Here, $V_{max,\dots}$ are the unknown maximum reaction speeds, $K_{m,\dots}$ the unknown Michaelis-Menten constants, and $c_{\dots}$ are the measured concentrations in the experimental time series. The parameters $V_{max,\dots}$ include the known initial concentration of the enzyme c_{E0} such that corrected reaction speeds are given by

$$V_{max,corr} = \frac{V_{max,\dots}}{c_{E0}}.$$

Thus, the unknown parameter vector of both models is

$$\theta = (V_{max,BABZ}; V_{max,BZDHPP}; K_{m,BABZ}; K_{m,BZDHPP}) \in \mathbb{R}^4. \quad (9)$$

The measurements of the concentrations c_{BA}, c_{BZ} and c_{DHPP} at different time points determine that the observations are multivariate. It is assumed that measurement errors are the only cause for randomness. Since the equations 1 to 3 and 4 to 6, respectively, have no explicit solutions, both models are multivariate implicitly given nonlinear models.

In Section 2, we first discuss how an unknown parameter vector θ of a multivariate implicitly given nonlinear model can be estimated. We compare there four different estimates: an estimate based on least squares, a trimmed least squares estimator, a Monte Carlo estimator and an MC-estimator with extra-noise. In Section 3, these methods are then applied to the estimation of the parameter vector θ given at (9) from the measured concentrations c_{BA}, c_{BZ} and c_{DHPP}. Based on the parameter estimates, a discrimination between Model A and B is possible and optimal designs for parameter estimation as well as for the discrimination between several models can be derived. The problem of model discrimination is treated in Section 4 and design considerations are presented in Section 5. In this study not only the time points but also the initial concentrations of the enzyme BAL and the initial substrate BA are chosen appropriately for the design.

2 Parameter estimation in implicitly given nonlinear models

Assume that a multivariate observation $y_n = (y_{n1}, \ldots, y_{nI})$ at a time point t_n is given by

$$y_n = (y_{n1}, \ldots, y_{nI}) = (g_1(t_n, \theta), \ldots, g_I(t_n, \theta)) + (\epsilon_{n1}, \ldots, \epsilon_{nI}),$$

for $n = 1, \ldots, N$, where $\theta \in \Theta \subset \mathbb{R}^L$ is an unknown parameter vector, $\epsilon_{n1}, \ldots, \epsilon_{nI}$ are measurement errors and $g(t, \theta) = (g_1(t, \theta), \ldots, g_I(t, \theta))$ is given by the I-dimensional nonlinear differential equation system

$$\begin{aligned}
\frac{d}{dt} g_1(t, \theta) &= h_1(g_1(t, \theta), \ldots, g_I(t, \theta); \theta) \\
\frac{d}{dt} g_2(t, \theta) &= h_2(g_1(t, \theta), \ldots, g_I(t, \theta); \theta) \\
&\vdots \\
\frac{d}{dt} g_I(t, \theta) &= h_I(g_1(t, \theta), \ldots, g_I(t, \theta); \theta)
\end{aligned}$$

with

$$g(t_0, \theta) = a_0,$$

where t_0 is the initial time point and a_0 the vector of initial conditions. The time points satisfy $t_0 < t_1 < t_2 < \ldots < t_N$. The differential equation system can be solved stepwise by the one-step method of Euler: Let $\tau_0 < \tau_1 < \ldots < \tau_S$ be equidistant time points in $[t_0, t_N]$ so that $\{t_0, t_1, \ldots, t_N\} \subset \{\tau_0, \ldots, \tau_S\}$. Taylor's theorem provides

$$\begin{aligned}
& \frac{d}{dt} g_i(\tau_j, \theta) = \frac{g_i(\tau_{j+1}, \theta) - g_i(\tau_j, \theta)}{\tau_{j+1} - \tau_j} + E_j \\
\iff \quad & g_i(\tau_{j+1}, \theta) = g_i(\tau_j, \theta) + (\tau_{j+1} - \tau_j) \frac{d}{dt} g_i(\tau_j, \theta) - (\tau_{j+1} - \tau_j)\, E_j,
\end{aligned}$$

where $E_j \to 0$ if $\tau_{j+1} - \tau_j \to 0$. There exist several one-step methods which differ mainly by stopping rules based on the magnitude of $(\tau_{j+1} - \tau_j)\, E_j$ (see e.g. [DP58], [SW95]). In this paper, we used the method of Dormand-Prince, an explicit Runge-Kutta method of fourth order, which is implemented in Matlab®.

Any solver for differential equations provides for each θ solutions

$$\begin{array}{cccc} \widetilde{g}_1(\tau_1,\theta), & \widetilde{g}_2(\tau_1,\theta), & \dots & \widetilde{g}_I(\tau_1,\theta), \\ \widetilde{g}_1(\tau_2,\theta), & \widetilde{g}_2(\tau_2,\theta), & \dots & \widetilde{g}_I(\tau_2,\theta), \\ \vdots & \vdots & & \vdots \\ \widetilde{g}_1(\tau_S,\theta), & \widetilde{g}_2(\tau_S,\theta), & \dots & \widetilde{g}_I(\tau_S,\theta). \end{array}$$

The determination of the solutions at the time points $t_1, \dots, t_N$ yields the candidate matrix

$$\widetilde{g}(\theta) = (\widetilde{g}_{ni}(\theta))_{n=1,\dots,N,i=1,\dots,I} = \begin{pmatrix} \widetilde{g}_1(t_1,\theta) & \widetilde{g}_2(t_1,\theta) & \dots & \widetilde{g}_I(t_1,\theta) \\ \widetilde{g}_1(t_2,\theta) & \widetilde{g}_2(t_2,\theta) & \dots & \widetilde{g}_I(t_2,\theta) \\ \vdots & \vdots & & \vdots \\ \widetilde{g}_1(t_N,\theta) & \widetilde{g}_2(t_N,\theta), & \dots & \widetilde{g}_I(t_N,\theta) \end{pmatrix}.$$

The parameter vector θ shall be chosen in the way that the difference between the candidate matrix $\widetilde{g}(\theta)$ and the matrix of observations

$$y = (y_{ni})_{n=1,\dots,N,i=1,\dots,I}$$

is as small as possible. The difference between the candidate matrix and the matrix of observations can be measured by a weighted sum of squares leading to the weighted least squares estimator.

Definition 1 *The weighted least squares estimator $\hat{\theta}_{LS}$ for multivariate implicitly defined nonlinear models is defined as*

$$\hat{\theta}_{LS} \;=\; arg\min_{\theta\in\Theta} \sum_{n=1}^{N}\sum_{i=1}^{I} w_{ni}(y_{ni} - \widetilde{g}_{ni}(\theta))^2,$$

where $(w_{ni})_{n=1,\dots,N,i=1,\dots,I}$ are given nonnegative weights.

The weighted least squares estimator can be calculated by the Matlab® function *lsqnonlin* which is based on the Levenberg-Marquardt algorithm ([SW03]). This algorithm is an extension of the Gauss-Newton algorithm and is in our case based on the vector of weighted residuals

$$\begin{aligned} r(\theta) &= (r_1(\theta), \dots, r_{N\cdot I}(\theta))^\top \\ &= (\sqrt{w_{11}}(\widetilde{g}_{11}(\theta) - y_{11}), \dots, \sqrt{w_{N1}}(\widetilde{g}_{N1}(\theta) - y_{N1}), \dots, \\ &\qquad \sqrt{w_{1I}}(\widetilde{g}_{1I}(\theta) - y_{1I}), \dots, \sqrt{w_{NI}}(\widetilde{g}_{NI}(\theta) - y_{NI}))^\top \end{aligned} \tag{10}$$

and its derivative at $\theta^{(q)}$

$$K(\theta^{(q)}) = \frac{\partial}{\partial\theta}\, r(\theta)\Big|_{\theta=\theta^{(q)}} \in \mathbb{R}^{N\cdot I\times L}. \tag{11}$$

Then the iteration step of the algorithm is given by

$$\theta^{(q+1)} = \theta^{(q)} + \delta^{(q)}$$

where $\delta^{(q)}$ is determined by

$$\delta^{(q)} = -\left(K(\theta^{(q)})^\top K(\theta^{(q)})\right)^{-1} K(\theta^{(q)})^\top r(\theta^{(q)})$$

in the Gauss-Newton algorithm and determined by

$$\delta^{(q)} = -\left(K(\theta^{(q)})^\top K(\theta^{(q)}) + \eta^{(q)}\right)^{-1} K(\theta^{(q)})^\top r(\theta^{(q)})$$

in the Levenberg-Marquardt algorithm, where

$$\eta^{(q)} = \arg\max_{\eta \in \mathbb{R}^{L\times L}} \left\| \left(K(\theta^{(q)})^\top K(\theta^{(q)}) + \eta\right)^{-1} K(\theta^{(q)})^\top r(\theta^{(q)}) \right\|.$$

Since $\widetilde{g}(\theta)$ is given only approximately, also the Jacobi matrix $K(\theta)$ was approximated by the Matlab® function *lsqnonlin* using finite differences.

To avoid that some few outliers have strong influence on the estimator, the trimmed least squares estimator (LTS) can be used as an alternative to the weighted least squares estimator (see e.g. [RL87] and [SR92]).

Definition 2 *The least trimmed squares estimator $\hat{\theta}_{LTS}$ for multivariate implicitly defined nonlinear models is defined as*

$$\hat{\theta}_{LTS} \quad = \quad arg\min_{\theta\in\Theta} \sum_{m=1}^{N\cdot I-k} r^u_{(m)}(\theta)^2,$$

where $|r^u_{(1)}(\theta)| \le |r^u_{(2)}(\theta)| \le \ldots \le |r^u_{(N\cdot I)}(\theta)|$ are the ordered unweighted residuals given by $r^u_1(\theta) = (\widetilde{g}_{11}(\theta) - y_{11}), \ldots, r^u_{N\cdot I}(\theta) = (\widetilde{g}_{NI}(\theta) - y_{NI})$.

Numerical problems occured when the trimming proportion k is too large. However, the estimator could be calculated for $k = 1$ and $k = 2$ for the BAL catalyzed reactions.

To study the variability of an estimator, the following two Monte Carlo estimators were used.

Definition 3 *The initial estimator is any estimator $\hat{\theta}$, for which the solution of the differential equations $\widetilde{g}(\hat{\theta})$ is calculated. Then $\mathcal{N}(0,1)$ distributed errors are added to the matrix $\widetilde{g}(\hat{\theta})$ leading to a new observation matrix y^1 and a new estimator $\hat{\theta}(y^1)$. This is repeated several times, say K times, and the mean of the estimators $\hat{\theta}(y^1), \ldots, \hat{\theta}(y^K)$ is taken as the Monte Carlo estimator $\hat{\theta}_{MCData}$. Compare [ET98].*

Definition 4 *The initial estimator is like in Definition 3 any estimator $\hat{\theta}$. Then noise is added in two steps. In the first step, normal distributed errors are added to the starting vector $\hat{\theta}$ leading to a vector $\hat{\theta}^{1E}$. Then $\mathcal{N}(0,1)$ distributed errors are added to the matrix $\widetilde{g}(\hat{\theta}^{1E})$ leading to a new observation matrix y^{1E} and a new estimator $\hat{\theta}(y^{1E})$. Both steps are repeated several times, say K times, and the mean of the estimators $\hat{\theta}(y^{1E}), \ldots, \hat{\theta}(y^{KE})$ is taken as the MC-estimator $\hat{\theta}_{SIMUL}$.*

3 Parameter estimation for the example-reaction

The data of three batch experiments were used with the initial concentrations of BA and BAL listed in Table 1. The concentrations of BA, BZ, and DHPP were measured at time points 0, 16, 21, 30, 45, 60, 90, 120, 150, 180, 210, 240, 300 minutes. Hence, there were $N = 13$ time points, each with $I = 3 \cdot 3$ measurements, resulting in $N \cdot I = 117$ overall measurements.

	BA [mM]	BAL [mg/ml]
Batch A	56.0	0.22
Batch B	31.5	0.29
Batch C	18.7	0.20

Table 1: Initial concentrations of the Batch experiments

The time point t=0 at the beginning of the experiment with fixed concentrations was included to get an easy estimate of the measurement error. For the weighted least squares estimator the weights

$$w_{ni} = (a \cdot y_{ni} + 0.1)^{-1} \quad \text{with} \quad a = \frac{\widehat{Var}(y_{max}) - 0.1}{y_{max}}$$

with

$$y_{max} = \max\{y_{ni};\ n = 1, \ldots, N,\ i = 1, \ldots, I\}$$

were used, where the estimated variance $\widehat{Var}(y_{max})$ of the maximal measured value was given by the experimenter. These weights ensure that small values y_{ni} near 0 are weighted mainly by the basic noise given by 0.1 and the importance of the variance increases with increasing measurement values.

The initial values for the Levenberg-Marquardt algorithm were chosen randomly within the set $[0, 1000]^4$. The algorithm was repeated several times in order to ensure that the global maximum was found although there is of course no proof that this is really the case.

The trimming number for the least trimmed squares estimator was set to $k = 2$. The starting estimator for both Monte Carlo estimators was the weighted least squares estimator $\hat{\theta}_{LS}$, and $K = 1000$ was used as number of repetitions.

All four parameter estimates are shown for both reaction models in Figure 3 and Figure 4.

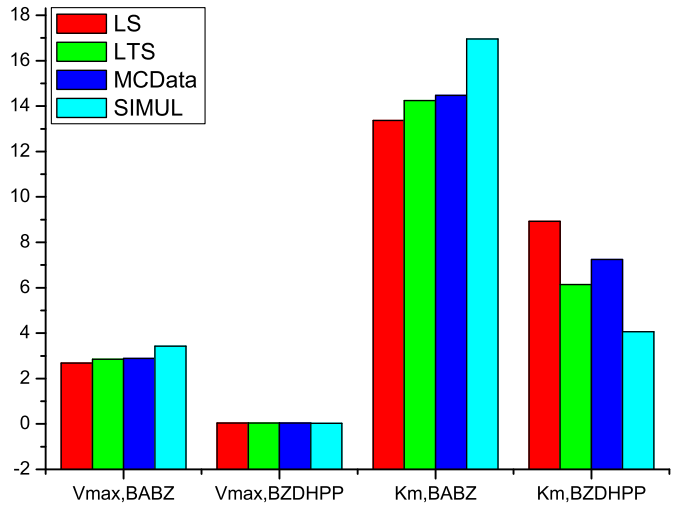

Figure 3: Parameter estimates for Reaction Model A

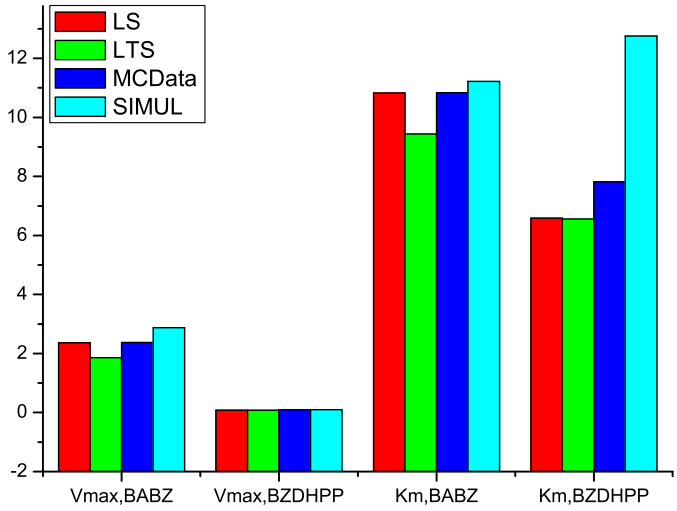

Figure 4: Parameter estimates for Reaction Model B

The four different parameter estimators show very similar behavior for both reaction models. The similarity of the weighted least squares estimator and the trimmed least squares estimator indicates that there are no heavy outliers. Only the estimates for $K_{m,BZDHPP}$ differ a little bit, a hint, that this parameter can not be determined as exactly as the others. An example for the fitting of the model to the data is shown in Figure 5 where Model B is fitted to the data from Batch A. All other fits look similar to this example.

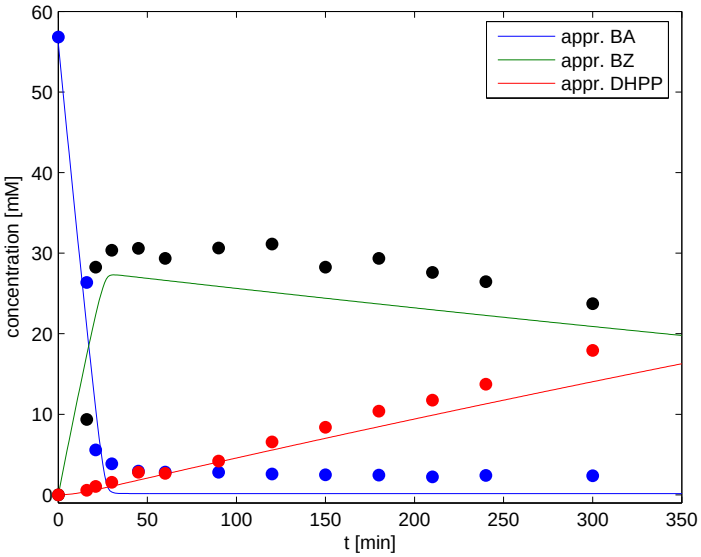

Figure 5: Model fit for Reaction Model B and the data from Batch A

4 Model discrimination

Model discrimination can be done with the Akaike information criterion which is defined as (see e.g. [FH95])

$$AIC = M \ \ln\left(\frac{\sum_{m=1}^{M} r_m(\hat{\theta})^2}{M}\right) + 2\ L,$$

where $M = N \cdot I$ and $r_1(\theta), \ldots, r_M(\theta)$ are the residuals defined in (10). For small samples the Small Sample AIC of [HT89] should be used

$$AIC_C = AIC + \frac{2L(L+1)}{M-L-1}.$$

For discrimination of the two reaction models, we have for both models $L = 4$ and a comparison with the sum of squared residuals could be done. However, for further model discrimination, Table 2 shows also the Small Sample AIC for the the weighted least squares estimator.

	Reaction Model A	Reaction Model B
$\sum_m r_m(\hat{\theta})^2$	779.32	532.72
AIC_C	229.86	185.35

Table 2: Comparison of the sum of squared residuals and the AIC_C values of $\hat{\theta}_{LS}$ for the both reaction models

Table 2 shows that Model B provides the smaller AIC and therefore the better model fit. To study whether this is not caused by measurement error, the simulated estimators $\hat{\theta}(y^1), \ldots, \hat{\theta}(y^{1000})$ for getting the Monte Carlo estimator $\hat{\theta}_{MCData}$ were used to get a distribution of the differences between the Akaike values for the both models. The results are shown in Figure 6.

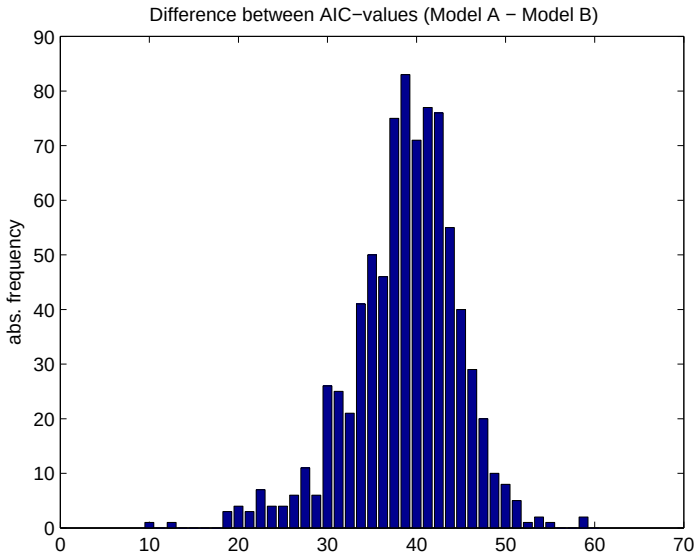

Figure 6: Distribution of the AIC_C differences (Model A - Model B)

Although the differences are always positive, i.e. Model B is for every MC-dataset 'better' than Model A, the question is whether a better discrimination between the two models is possible with another experimental design. This is the topic of the next section.

The Akaike criterion is needed to discriminate models with a different number of parameters, shown by the following example. A modification of Model B can be obtained by regarding Model C given by

$$\begin{aligned} \frac{d\ c_{BA}}{dt} &= -2r_1 + r_3 - r_5, \\ \frac{d\ c_{BZ}}{dt} &= r_1 - r_3, \\ \frac{d\ c_{DHPP}}{dt} &= r_3 + r_5, \end{aligned}$$

where

$$\begin{aligned} r_1 &= V_{max,BABZ} \cdot \left(\frac{c_{BA}}{K_{M,BABZ} + c_{BA}} \right)^2, \\ r_3 &= V_{max,BZDHPP} \cdot \frac{c_{BZ}}{K_{M,BZDHPP} + c_{BZ}}, \\ r_5 &= V_{max,BADHPP} \cdot \frac{c_{BA}}{K_{M,BADHPP} + c_{BA}}. \end{aligned}$$

Model C has $L = 6$ parameters. Here, the sum of squared residuals for the weighted least squares estimate is 527.27, so that we obtain $AIC_C = 188.15$. Hence Model C provides a larger AIC than Model B.

5 Experimental Designs

Optimal designs for parameter estimation minimize a functional of the covariance matrix of the parameter estimator. In nonlinear models, only an asymptotic or

approximate covariance matrix of the least squares estimator can be derived. This approximation depends on the unknown parameter vector θ and is given by (see [BB89], [Páz93], [SW03])

$$(J(\theta)^\top J(\theta))^{-1},$$

where $J(\theta)$ in the linearized system is defined as derivative of the true unknown non-linear function g with respect to θ (compare to (11)). As soon as there is no explicit expression for g, g must be replaced by the calculated solution of the differential equations $\widetilde{g}$ and $J(\theta)$ must be approximated by $\widetilde{J}(\theta)$ given by

$$\widetilde{J}(\theta)_{ml} = \frac{\widetilde{g}_m(\theta + h_l\ e_l) - \widetilde{g}_m(\theta - h_l\ e_l)}{2\,h_l} \tag{12}$$

for sufficient small $h_l \in (0,1)$, $m = 1, \ldots, N \cdot I$ and $l = 1, \ldots, L$, where e_l is the l'th unit vector of $\mathbb{R}^L$. Hence, the differential equations solver has to be applied $2 \cdot L$ times per considered parameter vector θ to provide $\widetilde{g}(\theta + h_l\ e_l)$ and $\widetilde{g}(\theta - h_l\ e_l)$. This is similar to the approximation used by Matlab® in the Levenberg-Marquardt algorithm. Here a self-implemented routine was applied and the calculation of the approximated Jacobi matrix $\widetilde{J}(\theta)$ took four to ten seconds for the models of the BAL catalyzed reactions.

To express the dependence of the design δ, we use here the notation $\widetilde{J}(\theta, \delta)$ instead of $\widetilde{J}(\theta)$. D- and E-optimal designs can be now defined analogously to the definitions in [Páz86] and [Páz93].

Definition 5
a) A design δ_* is called locally D-optimal in Δ, if it satisfies

$$\delta_* = \arg\min\{\det(\widetilde{J}(\theta, \delta)^\top \widetilde{J}(\theta, \delta))^{-1};\ \delta \in \Delta\}.$$

b) A design δ_* is called locally E-optimal in Δ, if it satisfies

$$\delta_* = \arg\min\{\lambda_{\max}(\widetilde{J}(\theta, \delta)^\top \widetilde{J}(\theta, \delta))^{-1};\ \delta \in \Delta\},$$

where $\lambda_{\max}$ denotes the maximum Eigen value.

The locally optimal designs depend on the unknown parameter vector θ and can only be used if an estimate of θ is available. Here, we used the weighted least squares estimators of Section 3 so that the proposed covariance matrix is given by $(\widetilde{J}(\hat{\theta}_{LS}, \delta)^\top \widetilde{J}(\hat{\theta}_{LS}, \delta))^{-1}$.

The experimental designs consist of the initial concentrations of BA and BAL and the time points $t_1, \ldots, t_N$. For parameter estimation, we consider only designs that differ in the initial concentrations of BA and BAL. The results for the D-criterion are given in Figure 7 and Figure 8. The plus signs are marking the standard deviation of the proposed values when the D-criterion is calculated for the simulated estimators $\hat{\theta}(y^1), \ldots, \hat{\theta}(y^{1000})$ from Definition 3.

Similar dependencies on the initial concentrations are obtained for the maximum Eigen values of the covariance matrices. Hence a D- and E-optimal design will be a design with minimum possible initial concentration of the enzyme BAL and with maximum possible concentration of the substrate BA.

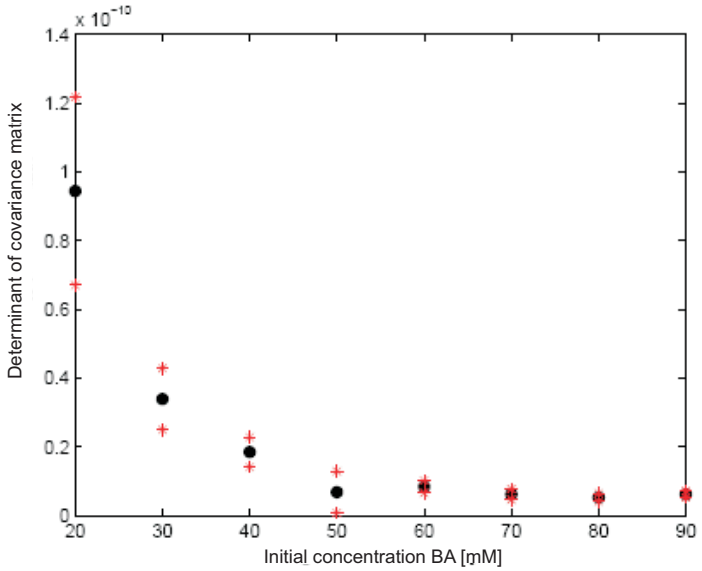

Figure 7: Dependency of the determinant of the covariance matrix on the initial concentration of the substrate BA

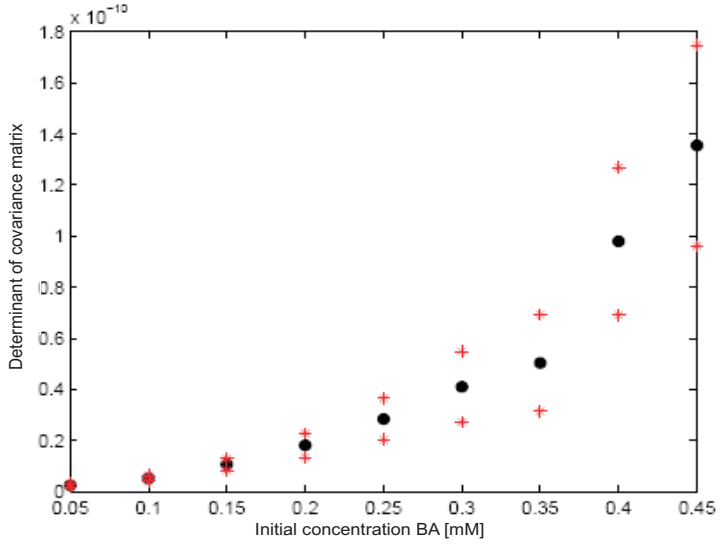

Figure 8: Dependency of the determinant of the covariance matrix on the initial concentration of the enzyme BAL

The same result holds for model discrimination between the Models A and B. For defining a criterion for model discrimination let

$$\widetilde{g}_A(\theta,\delta) = (\widetilde{g}_{Ani}(\theta,\delta))_{n=1,\ldots,N,i=1,\ldots,I} = (\widetilde{g}_i(t_n,\theta))_{n=1,\ldots,N,i=1,\ldots,I}$$

be the calculated solution for Model A for design δ, and $\widetilde{g}_B(\theta,\delta)$ is defined analogously. An optimal design for model discrimination should maximize the distance between $\widetilde{g}_A(\theta,\delta)$ and $\widetilde{g}_B(\theta,\delta)$ [AD92].

Definition 6 *A design δ_* is locally optimal in Δ for discrimination between the Models A and B if*

$$\delta_* = \arg\max\{\sum_{n=1}^{N}\sum_{i=1}^{I} |\widetilde{g}_{Ani}(\theta,\delta) - \widetilde{g}_{Bni}(\theta,\delta)|;\ \delta \in \Delta\}.$$

Again we use the weighted least squares estimator for θ. The dependencies of the distance between the models on the initial concentrations are given in Figure 9 and Figure 10.

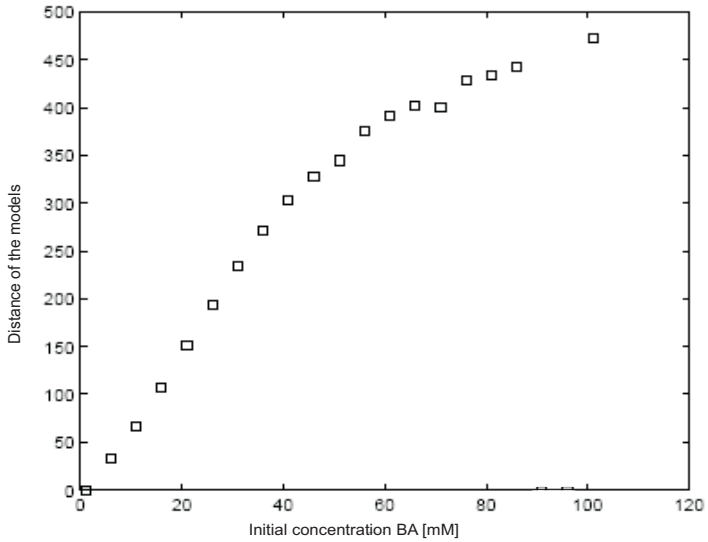

Figure 9: Dependencies of the distances between Model A and Model B on the initial concentration of substrate (BA)

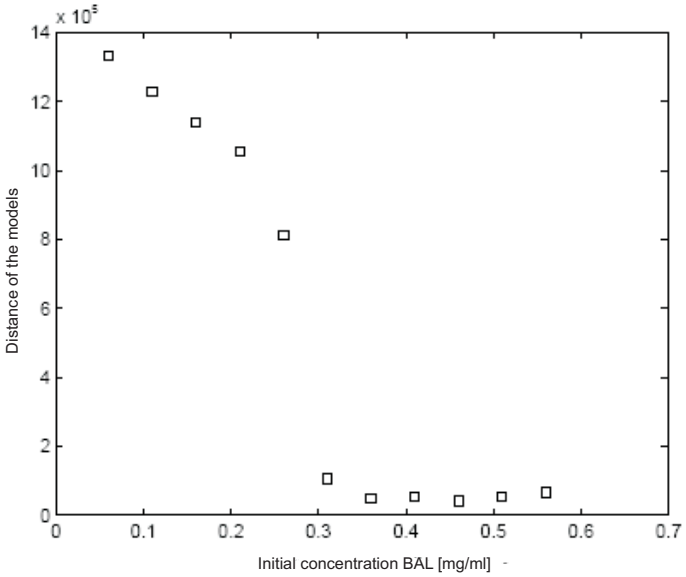

Figure 10: Dependencies of the distances between Model A and Model B on the initial concentration of the enzyme (BAL)

Hence an optimal design for parameter estimation for Model A and Model B is also an optimal design for discrimination between the two models.

To study also the influence of the time points $t_1, \ldots, t_N$ on the distance between the two models, the following designs were regarded

- Ä_300: Equidistant time points up to the 300'th minute
- 2:1_300: Time points up to the 300'th minute, whereof two third of the measurements are done at equidistant time points in the first half of the interval
- Ä_150: Equidistant time points up to the 150'th minute
- 2:1_150: Time points up to the 150'th minute, whereof two third of the measurements are done at equidistant time points in the first half of the interval
- Expo: Time points with distances which increase exponentially (2^x) up to 300'th minute

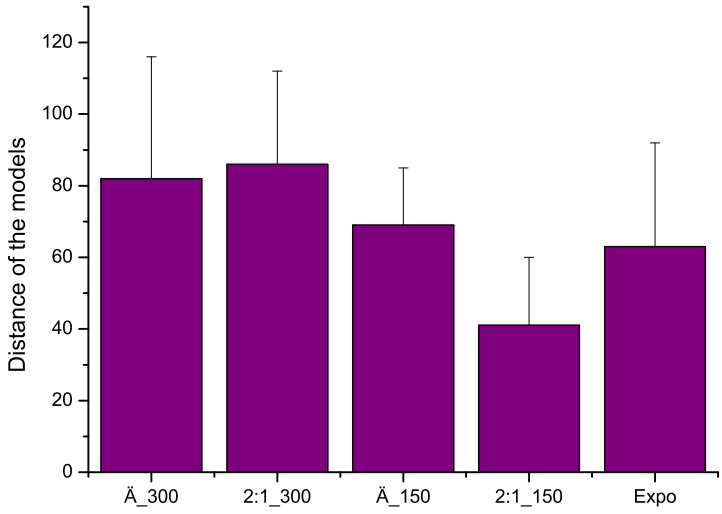

Figure 11: Comparison of different designs for the time points

Figure 11 shows that a good design is a design with equidistant points in the longest interval. However, the standard deviations given by the simulated estimators $\hat{\theta}(y^1), \ldots, \hat{\theta}(y^{1000})$ of Definition 3 are rather large and the other designs could consequently have a similar quality for discriminating between the two models.

Acknowledgement

We gratefully thank Sven Kühl and Stephan Lütz from Research Centre Jülich for kindly providing us with the experimental data and supporting us with their biochemical knowledge.

References

[AD92] A.C. Atkinson and A.N. Donev. Optimum experimental designs. In *Oxford Statistical Science Series*. Oxford University Press, Oxford, 1992.

[BB89] H. Bunke and O. Bunke. *Nonlinear Regression, Functional Relations and Robust Methods*. John Wiley & Sons, Inc., Berlin, 1989.

[DEH+02] A.S. Demir, E. Eren, B. Hosrik, Ö. Şeşenoglu, M. Pohl, E. Janzen, D. Kolter, R. Feldmann, P. Dünkelmann, and M. Müller. Enantioselective synthesis of α-hydroxy ketones via benzaldehyde lyase-catalyzed c-c bond formation reaction. *Advanced Synthesis & Catalysis*, 344(1):96 – 103, 2002.

[DP58] J.R. Dormand and P.J. Prince. A family of embedded runge-kutta formulae. *Journal of Computational Mathematics*, 10:517–534, 1958.

[ET98] B. Efron and R.J. Tibishirani. An introduction to the bootstrap. In *Monographs on Statistics and Applied Probability 57*. Chapman & Hall/CRC, Boca Raton, 1998.

[FH95] L. Fahrmeir and A. Hamerle. *Multivariate statistische Verfahren*. Walter de Gruyter & Co., Berlin, 1995.

[HT89] C.M. Hurvich and C-L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307, 1989.

[Küh07] S. Kühl. *Enzymkatalysierte C-C Knüpfung: Reaktionstechnische Untersuchungen zur Synthese pharmazeutischer Intermediate*. PhD thesis, Universität Bonn, 2007.

[Páz86] A. Pázman. *Foundations of Optimum Experimental Design*. Reidel, Dordrecht, 1986.

[Páz93] A. Pázman. *Nonlinear Statistical Models*. Kluwer, Dordrecht, 1993.

[RL87] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. Wiley, New York, 1987.

[SR92] A.J. Stromberg and D. Ruppert. Breakdown in nonlinear regression. *J. Amer. Statist. Assoc.*, 87:991–997, 1992.

[SW95] K. Strehmel and R. Weiner. *Numerik gewöhnlicher Differentialgleichungen*. B.G. Teubner, Stuttgart, 1995.

[SW03] G.A.F. Seber and C.J. Wild. *Nonlinear Regression*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2003.