A new model for kinetic parameter estimation in biochemical reactions

Paola Lecca
The Microsoft Research - University of Trento
Centre for Computational and Systems Biology
lecca@cosbi.eu

Guido Sanguinetti
Dept. of Computer Science, University of Sheffield, UK
guido@dcs.shef.ac.uk

Alida Palmisano
The Microsoft Research - University of Trento
Centre for Computational and Systems Biology
DISI - University of Trento
palmisano@cosbi.eu

Corrado Priami
The Microsoft Research - University of Trento
Centre for Computational and Systems Biology
DISI - University of Trento
priami@cosbi.eu
A new model for kinetic parameter estimation in biochemical reactions.

Paola Lecca, Guido Sanguinetti, Alida Palmisano, and Corrado Priami

Abstract

We present a novel method for estimating rate coefficients from noisy observations of concentration levels at discrete time points. This is traditionally done by computing the least-squares estimator. However, estimation of the error function generally requires solving the reaction rate equations, which can become computationally unfeasible. Here we present an alternative approach based on a probabilistic, generative model of the variations in reactant concentration. Our method returns the rate coefficients, the level of noise and an error range on the estimates of rate constants. Its probabilistic formulation is key to a principled handling of the noise inherent in biological data, and it allows a number of further extensions.
1 Introduction

The ability to infer kinetic parameters of biochemical reactions is emerging as a crucial problem in system biology. Estimates of reaction rates are often crucial for modelling and simulating the biochemical dynamics, yet their direct measurement is an outstanding experimental problem. The current need for new techniques for integrative and predictive modelling \[1, 7, 6\] is based on the realisation that traditional reductionist approaches have to be complemented by the reconstruction of a model of how systems function as a whole. The importance and the great interest that the scientific community attributes to this problem are evidenced by the increasing number of recent works concerning this research topic.

In this paper we present a novel method for estimating rate coefficients from noisy observations of concentration levels at discrete time points. Given a number of \(N\) reactant species, we observe time series concentrations for each of the species, gathered in \(N\) state vectors \(\vec{X}_1, \ldots, \vec{X}_N\). The relation between the instantaneous rate of reaction and the concentrations of the reactants at any moment is given by the law of mass action:

\[
\text{i.e. the rate at which a substance takes part in a reaction is proportional to its concentration raised to a power which represents the number of molecules taking part in the reaction. Such formulation is made for simultaneous as well as isolated reactions, and for heterogeneous as well as homogeneous systems. The goal is the estimation of the constant of proportionality. This is traditionally done by computing the least-squares estimators (LSE), resulting from the minimisation of the sum of squared differences between the observed data and the model. However, estimation of the error function generally requires solving the reaction rate equations, which can easily become computationally unfeasible. Here we present an alternative approach based on a probabilistic, generative model of the variations in reactant concentration.}
\]

Our method discretizes the law of mass action and provides a tool to predict the values of the variables \(\vec{X}_i\) at time \(t\), conditioned on their values at the previous time point. The variations of the species concentration at different time points are conditionally independent by the Markovian nature of the discrete model of the law of mass action. Assuming the observation noise to be Gaussian with variance \(\sigma^2\), the probability of observing a variation \(D_i\) for the concentration \([X]_i\) of species \(i\) between time \(t_{k-1}\) and \(t_k\) is a Gaussian with variance depending on \(\sigma\) and mean the expectation value of the law mass action function under the noise distribution. The discretization of the law of mass action provides a model for the variations of the species concentration, rather than a model for the time-trajectory of the species concentrations. This makes the evaluation of the expectation value of law mass action function simpler and analytically tractable. The rate coefficients and the level of noise \(\sigma^2\) are then obtained by maximising the likelihood function defined by the observed variations. Our method returns
the rate coefficients, the level of noise and an error range on the estimates of rate constants. Its probabilistic formulation is key to a principled handling of the noise inherent in biological data, and it allows a number of further extensions, such as a fully Bayesian treatment of the parameter inference and automated model selection strategies based on the comparison between marginal likelihoods of different models. Finally, the implementation of this method may be used as an interface tool, connecting the outcomes of the wet-lab activity for the concentration measurements and the softwares for the simulation of chemical kinetics. We tested our algorithm on some real case-study, whose rate coefficients were experimentally determined and well documented in literature for simple chemical reactions and for more complex biochemical interactions.

2 The model for inference

Consider $N$ reactant species, $S_1, S_2, \ldots, S_N$, with concentrations $X_1, X_2, \ldots, X_N$, that evolve according to a system of rate equations

$$\frac{dX_i}{dt} = f_i(X^{(i)}(t); \theta_i)$$

where $\theta_i$, $i = 1, 2, \ldots, N$, is the vector of the rate coefficients, which are present in the expression of the function $f_i$. We wish to estimate the set of parameters $\Theta = \cup \theta_i$ ($i = 1, 2, \ldots, N$), whose element $\theta_i$ is the set of rate coefficients appearing in the rate equations of i-th species, therefore

$$\theta_1 = \{\theta_{11}, \theta_{12}, \ldots, \theta_{1N_1}\}, \ldots, \theta_N = \{\theta_{N1}, \theta_{N2}, \ldots, \theta_{NN}\}$$

$X^{(i)}$ is the vector of concentrations of chemicals that are present in the expression of the function $f_i$ for the species $i$. According to the law of mass action, the functions $f_i$ have the general form

$$f_i(X^{(i)}(t); \theta_i) =$$

$$= \theta_{i1} \prod_{w \in S_1 \subseteq [1,N]} X_w^{\alpha_w} + \ldots + \theta_{iN_i} \prod_{w \in S_{N_i} \subseteq [1,N]} X_w^{\alpha_w} = \sum_{h=1}^{N_i} \left( \theta_{ih} \prod_{w \in S_h} X_w^{\alpha_w} \right)$$

where $\alpha_w \in \mathbb{R}$, and $N_i$ is the number of parameter in the $f_i$ rate equation. The rate equations in (2) form the so-called Generalized Mass Action law. We assume we have noisy observations $\hat{X}_i = X_i + \epsilon$ at times $t_0, \ldots, t_M$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise term with mean zero and variance $\sigma$. With this choice we are assuming that the concentration measurements are not significantly affected by systematic errors, but by uncontrolled random
errors and that an error is equally likely to occur in either positive or negative
direction with respect to the symmetry axis of the distribution.

We also assume a number \( M \) of concentration measurements for each
considered species. Approximating the rate equation (1) as a finite difference
equation between the observation times, gives

\[
X_i(t_k) = X_i(t_{k-1}) + (t_k - t_{k-1}) f_i(X^{(i)}(t_{k-1}); \theta_i)
\]

where \( k = 1, \ldots, M \). In Eq. (3) the rate equation is viewed as a model
of increments/decrements of reactant concentrations; i.e., given a value of
the variables at time \( t_{k-1} \), the model can be used to predict the value
at the next time point \( t_k \). Increments/decrements between different time
points are conditionally independent by the Markov nature of the model
(3). Therefore, given the Gaussian model for the noise, it is possible to
estimate the probability to observe the value \( \hat{X}_i(t_k) \) given the model at time
\( t_{k-1} \), \( X_i(t_{k-1}) \), and the set of parameters \( \theta_i \), as

\[
p\left( \hat{X}_i(t_k) | X_i(t_{k-1}) \right) = \mathcal{N} \left( X_i(t_{k-1}) + (t_k - t_{k-1}) f_i(X_i(t_{k-1}, \theta_i)), \sigma^2 \right)
\]

(4)

We then also have that the true value of \( X_i(t_k) \) is normally distributed
around the observed value \( \hat{X}_i(t_k) \), so that

\[
p\left( X_i(t_k) | \hat{X}_i(t_{k-1}) \right) = \mathcal{N} \left( \hat{X}_i(t_{k-1}), \sigma^2 \right) =
\]

\[
= \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{(X_i(t_k) - \hat{X}_i(t_{k-1}))^2}{2\sigma^2}\right]
\]

Therefore, the probability to observe a variation \( D_i(t_k) = X_i(t_k) - X_i(t_{k-1}) \)
for the concentration of the \( i \)-th species between the time \( t_{k-1} \) and \( t_k \), given
the parameter vector \( \theta_i \) is

\[
p(D_i(t_k)|\theta_i, \sigma) = \mathcal{N} \left( E[f_i(X^{(i)}(t_{k-1}), \theta_i)], 2\sigma^2 \right)
\]

(6)

and

\[
E[f_i(X^{(i)}(t_{k-1}, \theta_i))] = \int_{\Omega_X^{(i)}} f_i(X^{(i)}(t_{k-1}), \theta_i) \prod_{i=1}^{K_i} \left[ p_i \left( X_i(t_{k-1}) | \hat{X}_i(t_{k-1}) \right) \right] dX^{(i)}
\]

(7)

where \( \Omega_X^{(i)} \) is the sample space of \( X^{(i)} \), and \( K_i \) is the number of chemical
species in the expression for \( f_i \). While the increments/decrements are conditionally
independent given the starting point \( X_i(t_k) \), the random variables
\( D_i(t_k) \) are not independent of each other. Intuitively, if \( X_i(t_k) \) happens to be
below its expected value because of random fluctuations, then the following
increment \( D_i(t_{k+1}) \) can be expected to be bigger as a result, while the
previous one $D_i(t_k)$ will be smaller. A simple calculation allows us to obtain the covariance matrix of the vector of increments for the $i$-th species. This is a banded matrix $C_i \equiv C = \text{Cov}(D_i)$ with diagonal elements given by

$$E[D_i^2(t_k) - E[D_i^2(t_k)] = 2\sigma^2$$

and a non-zero band above and below the diagonal given by

$$E[(D_i(t_k) - E[D_i(t_k)])(D_i(t_{k-1}) - E[D_i(t_{k-1})])] = -\sigma^2$$

with all other entries zero. The likelihood for the observed increments/decrements therefore will be

$$p(D|\Theta) = \prod_{i=1}^{N} N(D_i|m_i(\Theta), C) = \left( \frac{1}{\sqrt{2\pi \det(C)}} \right)^{N} e^{-\frac{1}{2} \sum_{i=1}^{N} (D_i - m_i)^T C^{-1} (D_i - m_i)}$$

where $D = \{D_1, \ldots, D_N\}$, $D_i = D_i(t_1), D_i(t_2), \ldots D_i(t_M)$ ($i = 1, 2, \ldots, N$), and $m_i(t_{k-1}) \equiv E[f_i(X(t_{k-1}), \theta_i)]$.

The Eq. (8) can be optimized w. r. t. the parameters $\Theta = (\theta_1, \theta_2, \ldots, \theta_N)$ of the model to yield estimates of the parameters themselves and of the noise level. The chief numerical problem of this approach is the computation of the expectations of the rate functions given by equation (7). Non-integer values of the coefficients $\alpha$ can make estimating the integral analytically difficult. We propose an approximate method in which the Gaussian noise is replaced by an approximate uniform (white) noise, with the amplitude of the uniform noise being obtained as a sample from the Gaussian cumulative distribution function. At the first order, for small $\sigma$, we can approximate the Gaussian with zero mean and variance $\sigma$ with an uniform distribution defined on the interval $[-\sqrt{2\pi\sigma}/4, \sqrt{2\pi\sigma}/4]$, so that

$$\prod_{i=1}^{K_i} p_i = \prod_{i=1}^{K_i} \chi_i$$

where

$$\chi_i(X_i) = \begin{cases} \frac{2}{\sqrt{2\pi\sigma}} & \text{if } -\sqrt{2\pi\sigma}/4 \leq X_i \leq \sqrt{2\pi\sigma}/4 \\ 0 & \text{otherwise}. \end{cases}$$

This approximation makes the calculation of the expectation value of the rate equation (Eq. (7)) simpler and reduces the computational time of the procedure. Moreover, experiments not illustrated in this paper demonstrate that it does not influence the accuracy of the parameter estimates until $\sigma$ is less that 30% of the concentration measurement.

Substituting Eq. (9) in Eq. (7) gives
\[ E[f_i(X^{(i)}(t_{k-1}), \theta)] = \left( \frac{2}{\sqrt{2\pi} \sigma} \right)^{K_i} \int_{X_1-\sqrt{\frac{2\pi}{4}}}^{X_1+\sqrt{\frac{2\pi}{4}}} f_i(X^{(i)}(t_{k-1}), \theta_i) dX^{(i)} \] (10)

Now, substituting Eq. (2) in Eq. (10) leads to

\[ E[f_i(X^{(i)}(t_{k-1}), \theta_i)] = \left( \frac{2}{\sqrt{2\pi} \sigma} \right)^{K_i} \sum_{h=1}^{N} \theta_{ih} \left( \left( \frac{\sqrt{2\pi} \sigma}{2} \right)^{(S-S_{h})} \times \prod_{w \in S_{h}} \frac{1}{\alpha_{w}+1} \left( (\hat{X}_w + \sqrt{\frac{2\pi}{4}}) \alpha_{w}+1 - (\hat{X}_w - \sqrt{\frac{2\pi}{4}}) \alpha_{w}+1 \right) \right) \] (11)

where \( S \) is the set containing the indexes referring to all the \( K_i \) species appearing in \( f_i \), and \( \alpha_{w} \neq -1 \). In case some orders are equal to \(-1\) Eq. (11) takes the following form

\[ E[f_i(X^{(i)}(t_{k-1}), \theta_i)] = \left( \frac{2}{\sqrt{2\pi} \sigma} \right)^{K_i} \sum_{h=1}^{N} \theta_{ih} \left( \left( \frac{\sqrt{2\pi} \sigma}{2} \right)^{(S-S_{h})} \times \prod_{w \in S'_{h}} \frac{1}{\alpha_{w}+1} \left( (\hat{X}_w + \sqrt{\frac{2\pi}{4}}) \alpha_{w}+1 - (\hat{X}_w - \sqrt{\frac{2\pi}{4}}) \alpha_{w}+1 \right) \right) \times \prod_{w \in S''_{h}} \ln \frac{\hat{X}_w + \sqrt{\frac{2\pi}{4}}}{\hat{X}_w - \sqrt{\frac{2\pi}{4}}} \] (12)

where \( S'_{h} \) is the set of indexes \( \{h'_1, h'_2, \ldots, h'_s\} \) such that \( \alpha_{h'_l} \neq -1 \ \forall h'_l \in S'_{h} \), and \( S''_{h} \) is the set of indexes \( \{h''_1, h''_2, \ldots, h''_s\} \) such that \( \alpha_{h''_l} = -1 \ \forall h''_l \in S''_{h} \).

If in the Eq. (8), \( m_i \) is substituted with the expression (11) or (12), Eq. (8) becomes more tractable and can be optimized w. r. t. the parameters \( \Theta = (\theta_1, \theta_2, \ldots, \theta_N) \) and \( \sigma \). The values of the model’s parameters for which \( p(D|\Theta) \) has a maximum are the most likely values giving the observed kinetics.

### 3 The estimation errors

For a zero-th order reaction

\[
\begin{align*}
\emptyset & \xrightarrow{\theta} A \\
A & \xrightarrow{\theta} \emptyset
\end{align*}
\]
\[
\Delta \theta = \left\{ \left( \frac{\partial \theta}{\partial A_0} \epsilon_{A_0} \right)^2 + \left( \frac{\partial \theta}{\partial A_f} \epsilon_{A_f} \right)^2 \right\}^{\frac{1}{2}}
\]

(13)

\[
\frac{\partial k}{\partial A_f} = \frac{(A_f - A_0)}{t_f - t_0}
\]

(14)

\[
\frac{\partial k}{\partial A_0} = \frac{(A_0 - A_f)}{t_f - t_0}
\]

(15)

For a first order reaction

\[
A \overset{\theta}{\longrightarrow} P
\]

if \(A_0\) (\(A_f\)) is the initial (final) concentration of the molecules of the species \(A\), the error that the algorithm makes in estimating \(\theta\) is given by the following

\[
\Delta \theta = \left\{ \left( \frac{\partial \theta}{\partial A_0} \epsilon_{A_0} \right)^2 + \left( \frac{\partial \theta}{\partial A_f} \epsilon_{A_f} \right)^2 \right\}^{\frac{1}{2}}
\]

(16)

Since

\[
\theta = \frac{1}{t_f - t_0} \ln \left( \frac{A_0}{A_f} \right)
\]

(17)

Eq. 16 becomes

\[
\Delta \theta = \left\{ \left( \frac{\partial \theta}{\partial A_0} \epsilon_{A_0} \right)^2 + \left( \frac{\partial \theta}{\partial A_f} \epsilon_{A_f} \right)^2 \right\}^{\frac{1}{2}}
\]

(18)

For a second order reaction

\[
A + B \overset{\theta}{\longrightarrow} P
\]

if \(A_0\) (\(A_f\)) and \(B_0\) (\(B_f\)) are the initial (final) concentrations of the molecules of species \(A\) and \(B\), respectively, the error that the algorithm makes in estimating \(\theta\) is given by the following

\[
\Delta \theta = \left\{ \left( \frac{\partial \theta}{\partial A_0} \epsilon_{A_0} \right)^2 + \left( \frac{\partial \theta}{\partial B_0} \epsilon_{B_0} \right)^2 + \left( \frac{\partial \theta}{\partial A_f} \epsilon_{A_f} \right)^2 + \left( \frac{\partial \theta}{\partial B_f} \epsilon_{B_f} \right)^2 \right\}^{\frac{1}{2}}
\]

(19)

where

\[
\frac{\partial \theta}{\partial A_0} = \frac{1}{t_f - t_0} \left\{ \frac{1}{(A_0 - B_0)^2} \ln \frac{A_f B_0}{A_0 B_f} - \frac{1}{A_0 - B_0} \right\}
\]

(20)

\[
\frac{\partial \theta}{\partial B_0} = \frac{1}{t_f - t_0} \left\{ - \frac{1}{(A_0 - B_0)^2} \ln \frac{A_f B_0}{A_0 B_f} + \frac{1}{A_0 - B_0} \right\}
\]

(21)
\[
\frac{\partial \theta}{\partial A_f} = \frac{1}{t_f - t_0} \frac{1}{A_0 - B_0 A_f}, \quad (22)
\]
\[
\frac{\partial \theta}{\partial B_f} = \frac{1}{t_f - t_0} \frac{1}{B_0 - A_0 B_f}, \quad (23)
\]

4 Case studies

All the mathematical operations described above has been automated through the implementation of an algorithm in C++. We tested our algorithm on some real case-study, whose rate coefficients were experimentally determined and well documented in literature for simple chemical reactions and for more complex biochemical interactions. We show the ability of our algorithm of obtaining reasonable estimates for the rate coefficients in the following cases (see Table 1):

- **R1**. Bromine molecule formation $2Br \rightarrow Br_2$. The data referred to this reaction are the bromine concentration time-points after a flash photolysis of a mixture of bromine and $SF_6$ with $[Br_2]/[SF_6] = 3.2 \times 10^{-2}$ (experiment of Graff and Lang [11]).

- **R2**. In presence of an acid solution of phenol, the ion $IO_3^-$ is reduced to $IO_2^-$ by $Br^-$ accordingly to the reaction

  \[
  IO_3^- + 2Br^- + 2H^+ \rightarrow IO_2^- + Br_2 + H_2O
  \]

  With $[IO_3^-] = 5 \times 10^{-3} M$ and $[Br^-] = 1 \times 10^{-2} M$, Sharma and Gupta obtained the following data for a solution at 35 C with $[C_6H_5OH] = 2 \times 10^{-2} M$ [11].

- **R3**. the first-order decomposition in azomethane $CH_3N_2CH_3 \rightarrow CH_3CH_3 + N_2$ ([13]);

- **R4**. the alkaline hydrolysis of ethyl nitrobenzoate ([13]);

- **R5**. the glutathione S-transferase-catalysed dehalogenation of haloaromatic compounds with aromatic substrates ([15, 10]);

- **R6**. degradation of IkBa-UB in breast cancer cells (refer to [3] for a comparison);

- **R7**. Nuclear factor (NF) kB translocation from cytoplasm into the nucleus in breast cancer cells (r. t. [3] for a comparison);

- **R8**. the glucose consumption rate in astrocytes [9].
<table>
<thead>
<tr>
<th>Reaction</th>
<th>$k$ (experimental)</th>
<th>$k$ (estimate)</th>
<th>$\sigma$ (estimate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_1$</td>
<td>$(2.75 \pm 1.00) \times 10^9 M^{-1} s^{-1}$</td>
<td>$(1.93 \pm 1.28) \times 10^9 M^{-1} s^{-1}$</td>
<td>$3.2 \times 10^{-6}$</td>
</tr>
<tr>
<td>$R_2$</td>
<td>$(13.8 \pm 1.00) M^{-2} s^{-2}$</td>
<td>$(15.92 \pm 6.98) M^{-2} s^{-1}$</td>
<td>$3.5 \times 10^{-1}$</td>
</tr>
<tr>
<td>$R_3$</td>
<td>$(3.60 \pm 2.16) \times 10^{-4} s^{-1}$</td>
<td>$(0.80 \pm 0.45) \times 10^{-4} s^{-1}$</td>
<td>$5.36 \times 10^{-3}$</td>
</tr>
<tr>
<td>$R_4$</td>
<td>$(8.1 \pm 2.31) \times 10^{-2} s^{-1}$</td>
<td>$(5.9 \pm 1.5) \times 10^{-2} s^{-1}$</td>
<td>$3.3 \times 10^{-3}$</td>
</tr>
<tr>
<td>$R_5$</td>
<td>$(7.4 \pm 0.5) \times 10^{-4} s^{-1}$</td>
<td>$(2.60 \pm 1.17) \times 10^{-4} s^{-1}$</td>
<td>$4.1 \times 10^{-3}$</td>
</tr>
<tr>
<td>$R_6$</td>
<td>$(5.56 \pm 0.11) \times 10^{-4} s^{-1}$</td>
<td>$(3.1 \pm 2.2) \times 10^{-4} s^{-1}$</td>
<td>$5 \times 10^{-3}$</td>
</tr>
<tr>
<td>$R_7$</td>
<td>$(3.85 \pm 2.67) \times 10^{-4} s^{-1}$</td>
<td>$(1.3 \pm 0.4) \times 10^{-4} s^{-1}$</td>
<td>$8.5 \times 10^{-3}$</td>
</tr>
<tr>
<td>$R_8$</td>
<td>$(1.33 \pm 0.04) \times 10^{-2} s^{-1}$</td>
<td>$(1.24 \pm 0.78) \times 10^{-2} s^{-1}$</td>
<td>$6 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

Table 1: The estimates of the rate coefficients compared with the experimental values.

## 5 Conclusion

The ability of this algorithm to estimate the expected value of the kinetic rate constant significantly depends on the experimental error on the data. If the time-course measurements of the reactant concentration are affected by a high level of noise, this technique estimate a rate constant very close to zero, indicating that there is no dynamic in the systems. Therefore the observed variations in the concentration values are attributed exclusively to the noise. For an accurate estimate of the rate constants the procedure requires high-quality and high-resolution data, that at the time of writing are relatively difficult to obtain. In the meantime experimental biology evolves to make such data more readily available, the development of inference techniques allow identification of experimental design issues that need to be addressed to get more accurate data for more reliable estimate of rate constants.

## References


