A new method for inferring rate coefficients from experimental time-consecutive measurements of reactant concentrations

Paola Lecca^{1*}, Guido Sanguinetti^{2**}, Alida Palmisano^{1*}, Corrado Priami^{1*}
1. The Microsoft Research - University of Trento,
Centre for Computational and System Biology, Trento, Italy
2. Dept. of Computer Science, University of Sheffield, Sheffield, UK
*E-mail: {lecca, palmisano, priami}@cosbi.eu
**E-mail: G.Sanguinetti@dcs.shef.ac.uk

1 The model

We present a probabilistic, generative model of the *variations* in reactant concentration to infer the rate coefficients of a reaction from concentrations measured at discrete time point.

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*

$$\frac{dX_i}{dt} = f_i(X_1(t), X_2(t), \ldots, X_N(t); \Theta)), \tag{1}$$

where Θ is the vector of the rate coefficients $(\theta_1, \theta_2, \ldots, \theta_S)$ that we wish to estimate and f_i are the *rate functions* that have, according to the law of mass action, the general form

$$f_i(t) = \theta_1 g_1(\mathbf{X}(t)) + \theta_2 g_2(\mathbf{X}(t)) + \ldots + \theta_{N_i} g_{N_i}(\mathbf{X}(t)). \tag{2}$$

where $N_i \leq N$ and the function $g_i(\mathbf{X})$ is a product of reactant concentrations, so that

$$f_i(t) = \theta_1 \prod_{j \in S_1 \subseteq [1,N]} X_j^{\alpha_j} + \theta_2 \prod_{j \in S_2 \subseteq [1,N]} X_j^{\alpha_j} + \ldots + \theta_{N_i} \prod_{j \in S_{N_i} \subseteq [1,N]} X_j^{\alpha_j} \tag{3}$$

where $\alpha \in \mathbf{R}$. 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 σ^2. We discretize the rate equation (1) as a finite difference equation between the observation times,

$$X_i(t_k) = X_i(t_{k-1}) + (t_k - t_{k-1}) f_i(X_1(t_{k-1}), X_2(t_{k-1}), \ldots, X_N(t_{k-1}), \Theta) \tag{4}$$

where $k = 1, \ldots, M$. In Eq. (4) 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, the model can be used to predict the value at the next time point. Increments/decrements between different time points are conditionally independent by the Markovian nature of the model (4). 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 $\Theta = (\theta_1, \theta_2, \ldots, \theta_S)$, as

$$p\Big(\hat{X}_i(t_k) | X_i(t_{k-1}), \Theta\Big) = \mathcal{N}\Big(X_i(t_{k-1}) + f_i(\mathbf{X}(t_{k-1}), \Theta), \sigma^2\Big) \tag{5}$$

where $\mathbf{X}(t) = (X_1, X_2, \ldots, X_N)$. Moreover, by symmetry, the true value of $X_i(t_k)$ is normally distributed around the observed value $\hat{X}_i(t_k)$, so that

$$p\Big(X_i(t_k)|\hat{X}_i(t_k)\Big) = \mathcal{N}\Big(X(t_{k-1})|\hat{X}(t_{k-1}), \sigma^2\Big) = \frac{1}{\sqrt{2\pi}\sigma}\exp\Big[-\frac{(X_i(t_{k-1}) - \hat{X}_i(t_{k-1}))^2}{2\sigma^2}\Big] \quad (6)$$

Therefore, the probability to observe a variation $D_i(t_k)$ for the concentration of the i-th species between the time t_{k-1} and t_k, given the parameter set Θ is

$$p(D_i(t_k)|\Theta) = \mathcal{N}\Big(E[f_i(\mathbf{X}(t_{k-1}), \Theta)], 2\sigma^2\Big) \quad (7)$$

where $D_i(t_k) \equiv \hat{X}_i(t_k) - \hat{X}_i(t_{k-1})$ and

$$E[f_i(\mathbf{X}(t_{k-1}, \Theta))] = \int f_i(\mathbf{X}^{(i)}, \Theta) \prod_{i=1}^{K_i} \Big[p_i\Big(X_i(t_k)|\hat{X}_i(t_k)\Big)\Big] d\mathbf{X}^{(i)} \quad (8)$$

where $\mathbf{X}^{(i)}$ is the vector of chemicals appearing in the function f_i for the i-th species, and K_i is the number of chemical species present 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 $\mathbf{C}_i \equiv \mathbf{C} = \mathrm{Cov}(\mathbf{D}_i)$ with diagonal elements given by $E\Big[D_i^2(t_k) - E[D_i^2(t_k)]\Big] = 2\sigma^2$, and a non-zero band above and below the diagonal given by $E\Big[(D_i(t_k) - E[D_i(t_k)])(D_i(t_k) - E[D_i(t_{k-1})])\Big] = -\sigma^2$, with all other entries zero. The likelihood for the observed increments/decrements therefore will be

$$p(\mathbf{D}|\Theta) = \prod_{i=1}^N \mathcal{N}(\mathbf{D}_i|\mathbf{m}_i(\Theta), \mathbf{C}) = \Big(\frac{1}{\sqrt{2\pi\ \det(\mathbf{C})}}\Big)^N \exp\Big[\sum_{i=1}^N -\tfrac{1}{2}(\mathbf{D}_i - \mathbf{m}_i)^T \mathbf{C}^{-1}(\mathbf{D}_i - \mathbf{m}_i)\Big] \quad (9)$$

where $\mathbf{D} = \{\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_N\}$ and $\mathbf{m}_i(t_k) \equiv E\Big[f_i(\mathbf{X}(t_{k-1}), \Theta)\Big]$. The Eq. (9) can be optimized w. r. t. the parameters $\Theta = (\theta_1, \theta_2, \ldots, \theta_S)$ of the model to yield estimates of parameters themselves and of the noise level itself.

Sampling the experimental error ϵ and calculating $\prod_{i=1}^{K_i} p_i$. The chief numerical problem of the previous approach is the computation of the expectations of the rate functions given by equation (8). Non-integer values of the coefficients α 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.

As stated in the previous section the small causal experimental error ϵ follows a Gaussian distribution, i. e. $\epsilon \sim \mathcal{N}(0, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(\epsilon^2/(2\sigma^2)\Big)$. Therefore the cumulative distribution function (CDF) for ϵ is

$$\mathrm{CDF}(\epsilon) = \frac{1}{2}\Big(1 + \frac{2}{\sqrt{\pi}} + \int_0^{\frac{\epsilon'}{\sigma\sqrt{2}}} e^{-\epsilon^2} d\epsilon\Big) \quad (10)$$

Let r be a random number from an uniform distribution $\mathcal{U} \sim \mathcal{U}(0, 1)$; ϵ' can be calculated as the solution of the equation $CDF(\epsilon) = r$, that is $4\int_0^{\frac{\epsilon}{\sigma\sqrt{2}}} e^{-\epsilon'^2} d\epsilon' = \frac{\sqrt{\pi}}{2}\Big(r - \frac{1}{2}\Big)$. Here, the integral can be approximated by truncating at the first order the series expansion of the error function erf(ϵ), as ϵ is considered here a small casual error. Therefore we obtain $\epsilon \approx \frac{\sqrt{2\pi}}{2}\sigma\Big(r - \frac{1}{2}\Big)$. For each species i we generate a random number r_i and calculate an error ϵ_i according to this formula. Now, since $(X_i - \hat{X}_i)^2 = \epsilon^2$, from Eq. (6) we get

$$\prod_{i=1}^{K_i} p_i = \Big(\frac{1}{\sigma\sqrt{2\pi}}\Big)^{K_i} \cdot \exp\Big[-\sum_{i=1}^{K_i}\frac{\epsilon^2}{2\sigma^2}\Big] = \Big(\frac{1}{\sigma\sqrt{2\pi}}\Big)^{N_i} \exp\Big[-\sum_{i=1}^{N_i}\frac{\pi}{4}\Big(r_i - \frac{1}{2}\Big)^2\Big] \qquad (11)$$

Therefore

$$E[f_i(\mathbf{X}(t_{k-1}, \mathbf{\Theta}))] = \Big(\frac{1}{\sigma\sqrt{2\pi}}\Big)^{K_i} \cdot \exp\Big[-\frac{\pi}{4}\sum_{i=1}^{K_i}\Big(r_i - \frac{1}{2}\Big)^2\Big] \cdot \sum_{h=1}^{N_i}\theta_h \int \prod_{j\in S_h\subset[1,N]} X_j^{\alpha_j^{(h)}}\, d\mathbf{X}^{(i)} \qquad (12)$$

2 Case studies

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): R_1. the first-order decomposition in azomethane $CH_3N_2CH_3 \longrightarrow CH_3CH_3 + N_2$ ([1]); R_2. the alkaline hydrolysis of ethyl nitrobenzoathe ([1]); R_3. the gluthathione S-transferase-catalysed dehalogenation of haloaromatic compounds with aromatic substrates ([3]); R_4. degradation of IkBa-UB in breast cancer cells (refer to [4] for a comparison); R_5. Nuclear factor (NF) kB translocation from cytoplasm into the nucleus in breast cancer cells (r. t. [4] for a comparison); and R_6. the glucose consumption rate in astrocytes [2].

Reaction	k (sec^{-1}) (experimental)	k (sec^{-1}) (estimate)	σ (estimate)
R_1	$(3.60 \pm 2.16) \times 10^{-4}$	$(0.80 \pm 0.45) \times 10^{-4}$	5.36×10^{-3}
R_2	$(8.1 \pm 2.31) \times 10^{-2}$	$(5.9 \pm 1.5) \times 10^{-2}$	3.3×10^{-3}
R_3	$(7.4 \pm 0.5) \times 10^{-4}$	$(2.60 \pm 1.17) \times 10^{-4}$	4.1×10^{-3}
R_4	$(5.56 \pm 0.11) \times 10^{-4}$	$(3.1 \pm 2.2) \times 10^{-4}$	5×10^{-3}
R_5	$(3.85 \pm 2.67) \times 10^{-4}$	$(1.3 \pm 0.4) \times 10^{-4}$	8.5×10^{-3}
R_6	$(1.33 \pm 0.04) \times 10^{-2}$	$(1.24 \pm 0.78) \times 10^{-2}$	6×10^{-4}

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

References

[1] V. Nummert and M. Piirsalu, *Kinetic study of hydrolysis of benzoates*, J. Chem. Soc., Perkin Trans. **2**, pp. 583-594, 2000.

[2] P. Lecca and M. Lecca, *Molecular mechanism of glutamate-triggered brain glucose metabolism: a parametric model from FDG PET-scans*, In Procs of BVAI 2007, To appear in LNCS, 2007.

[3] B. Rossi, I. Mancini and G. Guella, *Kinetic studies of glutathione S-transferase-catalysed process through on- line liquid chromatography-electrospray mass spectrometry*, Philosophical Magazine, **84**(13-16), pp. 1373-1382, 2004.

[4] J. W. Fuseler, D. M. Merril, J. A. Rogers, M. B. Grisham and R. E. Wolf, *Analysis and quatitation of NF-kB nuclear translocation in tumor necrosis factor alpha (TNF-α) activated vascular endothelial cells*, Microsc. Microanal. **12**, pp. 269-276, 2006.