Model Parameter Estimation

Parameter Identifiability Theory

Parameter estimation for high-dimension, non-linear ODE models using dynamic data is an area of active research [Swameye et al., 2003; Polisetty et al., 2006; Rodriguez-Fernandez et al., 2006; Tucker et al., 2007], but it is currently unknown how to best deal with issues of parameter identifiability and uniqueness of the least squares optimum for models of dimensionality as high as ours. However, these issues are of primary importance in constructing such models, as lack of confidence in model parameters directly translates into lack of confidence in the model. In an effort to begin to address these important issues, we take the following approach to provide a measure of confidence in our parameter values.

The model consists of 233 parameters being fit to 378 experimental data points, so in principle the system is over-determined and the estimation problem is well-posed. In practice, however, given an m dimensional vector of parameter values $\boldsymbol{\Theta}$ and an n dimensional vector of experimental data $\mathbf{Y}$, only an m_I dimensional subset of the parameter vector $\boldsymbol{\Theta}_\mathbf{I}$ will be *identifiable*. Given $\mathbf{Y}$, only $\boldsymbol{\Theta}_\mathbf{I}$ can be estimated with reasonably high confidence, while the remaining $m_u = m - m_I$ dimensional vector of parameters $\boldsymbol{\Theta}_\mathbf{U}$ cannot be estimated from the experimental data with reasonable accuracy, and these parameters are called *unidentifiable*. $\boldsymbol{\Theta}_\mathbf{U}$ can be further divided into *structurally* and *practically* unidentifiable parameters, which we denote by the vectors $\boldsymbol{\Theta}_\mathbf{US}$ and $\boldsymbol{\Theta}_\mathbf{UI}$, respectively, and we elaborate on these distinctions later.

For a linear model, given a tolerance for parameter accuracy straightforward to determine $\boldsymbol{\Theta}_\mathbf{I}$ and $\boldsymbol{\Theta}_\mathbf{U}$, but for a non-linear model, without a detailed knowledge of the entire multi-dimensional parametric sensitivity space (computationally infeasible to obtain), it is impossible to determine which parameters are uniquely identifiable. However, it is possible to quantify *local identifiability*. Our modeling and parameter estimation problem is of the form

$$
\begin{aligned}
&\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, t, \boldsymbol{\Theta}) \\
&\mathbf{x}(t=0) = \mathbf{x}_\mathbf{0}(\boldsymbol{\Theta}) \\
&\hat{\mathbf{y}}(t, \boldsymbol{\Theta}) = \mathbf{g}(\mathbf{x}(t, \boldsymbol{\Theta})) \\
&\hat{\mathbf{Y}}(\mathbf{T}, \boldsymbol{\Theta}) = [\hat{\mathbf{y}}(t_1, \boldsymbol{\Theta}), \ldots, \hat{\mathbf{y}}(t_{n_t}, \boldsymbol{\Theta})]^T \\
&S(\mathbf{T}, \boldsymbol{\Theta}) = (\mathbf{Y}(\mathbf{T}) - \hat{\mathbf{Y}}(\mathbf{T}, \boldsymbol{\Theta}))^T W (\mathbf{Y}(\mathbf{T}) - \hat{\mathbf{Y}}(\mathbf{T}, \boldsymbol{\Theta}))
\end{aligned}
\qquad (1)
$$

where t is time, $\mathbf{x_o}$ are the initial conditions, $\mathbf{x}$ are the model state variables, $\mathbf{g}$ is a function relating the model state variables to the estimates of the experimental measurements at a particular time $\hat{y}$, $\hat{\mathbf{Y}}$ are model estimates of the experimental measurements $\mathbf{Y}$, $\mathbf{W}$ is a diagonal weighting matrix, $\mathbf{T}=(t_1, \ldots, t_{nt})$ is the n_t dimensional vector of the measurement time points and S is the sum of squared errors between model predictions and experimental data. It is desired to minimize S by changing $\mathbf{\Theta}$, so differentiating S with respect to $\mathbf{\Theta}$ and setting this equal to zero gives

$$\begin{aligned}
&\frac{dS}{d\mathbf{\Theta}} = \frac{d}{d\mathbf{\Theta}}\left((\mathbf{Y(T)} - \hat{\mathbf{Y}}(\mathbf{T},\mathbf{\Theta}))^T W (\mathbf{Y(T)} - \hat{\mathbf{Y}}(\mathbf{T},\mathbf{\Theta}))\right) = 0 \\
&\frac{d}{d\mathbf{\Theta}}\left(\mathbf{Y(T)}^T W \mathbf{Y(T)} - \mathbf{Y(T)}^T W \hat{\mathbf{Y}}(\mathbf{T},\mathbf{\Theta}) - \hat{\mathbf{Y}}(\mathbf{T},\mathbf{\Theta})^T W \mathbf{Y(T)} + \hat{\mathbf{Y}}(\mathbf{T},\mathbf{\Theta})^T W \hat{\mathbf{Y}}(\mathbf{T},\mathbf{\Theta})\right) = 0 \\
&\frac{d}{d\mathbf{\Theta}}\left(\mathbf{Y(T)}^T W \mathbf{Y(T)} - 2\cdot \mathbf{Y(T)}^T W \hat{\mathbf{Y}}(\mathbf{T},\mathbf{\Theta}) + \hat{\mathbf{Y}}(\mathbf{T},\mathbf{\Theta})^T W \hat{\mathbf{Y}}(\mathbf{T},\mathbf{\Theta})\right) = 0 \\
&-2\cdot \mathbf{Y}^T W \frac{d\hat{\mathbf{Y}}}{d\mathbf{\Theta}} + \hat{\mathbf{Y}}^T W \frac{d\hat{\mathbf{Y}}}{d\mathbf{\Theta}} + \frac{d\hat{\mathbf{Y}}}{d\mathbf{\Theta}} W \hat{\mathbf{Y}}^T = 0 \\
&\mathbf{Y}^T W \frac{d\hat{\mathbf{Y}}}{d\mathbf{\Theta}} - \mathbf{Y}^T W \frac{d\hat{\mathbf{Y}}}{d\mathbf{\Theta}} = 0 \\
&(\hat{\mathbf{Y}} - \mathbf{Y})^T W \frac{d\hat{\mathbf{Y}}}{d\mathbf{\Theta}} = 0 \Rightarrow \hat{\mathbf{Y}} = \mathbf{Y}
\end{aligned} \tag{2}$$

In a small neighborhood around the current parameter values $\mathbf{\Theta_c}$, $\hat{\mathbf{Y}}$ can be approximated using a 1st order Taylor series expansion

$$\begin{aligned}
&\hat{\mathbf{Y}}(\mathbf{T},\mathbf{\Theta}) = \hat{\mathbf{Y}}(\mathbf{T},\mathbf{\Theta_c}) + \left.\frac{\partial \hat{\mathbf{Y}}}{\partial \mathbf{\Theta}}\right|_{\mathbf{\Theta}=\mathbf{\Theta_c}} (\mathbf{\Theta} - \mathbf{\Theta_c}) \\
&\mathbf{\Delta\Theta} \equiv (\mathbf{\Theta} - \mathbf{\Theta_c}); \mathbf{Z} \equiv \left.\frac{\partial \hat{\mathbf{Y}}}{\partial \mathbf{\Theta}}\right|_{\mathbf{\Theta}=\mathbf{\Theta_c}} \\
&\hat{\mathbf{Y}}(\mathbf{T},\mathbf{\Theta}) = \hat{\mathbf{Y}}(\mathbf{T},\mathbf{\Theta_c}) + \mathbf{Z\Delta\Theta}
\end{aligned} \quad , \tag{3}$$

where $\mathbf{Z}$ is the numerically determined *n-by-m* parameter sensitivity matrix. Substituting (3) into the condition derived in (2) we obtain

$$\begin{aligned}
&\hat{\mathbf{Y}}(\mathbf{T},\mathbf{\Theta}) = \mathbf{Y(T)} = \hat{\mathbf{Y}}(\mathbf{T},\mathbf{\Theta_c}) + \mathbf{Z\Delta\Theta} \\
&\mathbf{\Delta Y} \equiv \mathbf{Y(T)} - \hat{\mathbf{Y}}(\mathbf{T},\mathbf{\Theta_c}) = \mathbf{Z\Delta\Theta} \\
&\mathbf{Z}^T \mathbf{\Delta Y} = \mathbf{Z}^T \mathbf{Z\Delta\Theta}
\end{aligned} \quad . \tag{4}$$

To solve for $\Delta\Theta$ it is necessary for the matrix $\mathbf{Z^TZ}$ to be of full rank so it can be inverted. Rank deficiency of $\mathbf{Z^TZ}$ means that there are structurally unidentifiable parameters, which we define as belonging to the vector Θ_{US}. The members of Θ_{US} correspond to the columns of the R matrix (obtained by performing QR decomposition—upper triangularization of $\mathbf{Z^TZ}$) which have no non-zero elements. The columns of $\mathbf{Z}$ corresponding to the parameters in Θ_{US} can then be removed to create a new, reduced sensitivity matrix $\mathbf{Z'}$, such that $\mathbf{Z'^TZ'}$ can be inverted.

Keeping the above consideration in mind and proceeding with the inversion we obtain

$$(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\Delta\mathbf{Y} = \Delta\boldsymbol{\Theta} \tag{5}$$

This is an analogous form to a multiple linear regression data fitting problem, where the *m-by-m* matrix $\mathbf{Z^TZ}$ is the Fisher Information Matrix. It follows then that in a small neighborhood around Θ_C, an approximate confidence interval for the i^{th} parameter is given by

$$ci_i = t_{\alpha/2}^{n-m}\sqrt{\frac{S}{n-m}}\sqrt{(\mathbf{Z^TZ})_{ii}^{-1}}\ , \tag{6}$$

where *ci* is the confidence interval and $t_{\alpha/2}^{n-m}$ is a t-distribution statistic calculated with *n-m* degrees of freedom at confidence level α. We define parameter *i* to be *practically locally identifiable* if

$$ci_i < \varepsilon_i\ , \tag{7}$$

where ε_i is a chosen tolerance. By specifying the tolerances for each parameter, Θ can be locally parsed into Θ_I and Θ_{UP}.

Parameter Estimation Procedure

We first constructed a set of initial parameter value guesses with reasonable upper and lower bounds (Supplementary Parameter Tables). These values and bounds came from previous models, *in vitro* experimental data, and physical limitations (e.g. diffusion limitations). Using these initial parameter values we performed local identifiability analysis as outlined above to find Θ_I, and then locally minimized *S* varying only Θ_I. The resulting parameter set did not yield a good fit to the data, but this was somewhat expected for a few reasons. First, there is no reason to believe that parameters from one model should be directly usable in another model, as unknown factors in the cellular

context surely affect the “effective” parameter values. For a similar reason there is no reason to believe that *in vitro* measurements of equilibrium constants, binding rates, and enzyme kinetic parameters will be directly usable in a model representing live cells. Lastly, the model is non-linear in the parameters and therefore the S surface has multiple local minima. Although there are many stochastic minimization algorithms that allow “escape” from local minima (e.g. Simulated Annealing and Genetic Algorithms), such algorithms are much more computationally expensive than local minimization and cannot guarantee finding a global optimum for such high-dimension parameter spaces. Therefore, we took the following approach:

1. Hand pick a subset Θ_v of the model parameters to vary. The parameters in this vector are chosen because we have little confidence in their initial values and/or are they have high sensitivity.
2. Pick random initial values for Θ_v from a uniform distribution between the upper and lower bounds.
3. Perform local minimization varying only Θ_v.
4. Repeat steps 2 and 3 to generate several candidate parameter sets for further refinement. Because our purpose here is generating several candidate parameter sets, we do not worry here about performing local identifiability analysis. This would constrain our search and possibly make us miss good parameter sets.

Using the above approach, we generated 137 candidate parameter sets, which took ~3 combined processor-years on several computers having varied hardware. To evaluate which parameter set(s) should be selected for further refinement, we made a histogram of S for all the candidates (Figure S5). From Figure S5, it is clear that one parameter set is far better than all the others (Normalized S =1), with the next best having double the S and most others having approximately an order of magnitude larger S. Visual inspection confirmed that this parameter set was a much better candidate for further refinement than the other low S sets.

Satisfied with this candidate parameter set, we proceeded to refine the entire parameter set employing the following strategy:

1. Perform local identifiability analysis to determine Θ_I and Θ_U at the current parameter values. We chose tolerances such that the approximate confidence intervals on identifiable parameters were ~±20%.
2. Implement a single iteration of a local minimization changing only Θ_I. Changing only Θ_I allows us to make the best use of our computational time.

3. Implement a single iteration of a local minimization varying the entire parameter set Θ to facilitate escaping a local minimum in Θ_I.
4. Repeat Steps 1 and 2 until S converges to a minimum.
5. If any parameter value lies on an upper or lower bound, expand this bound and repeat Steps 1-4.

The implementation of this procedure yields the final parameter estimates and locally identifiable parameters listed in Supplementary Parameter Table, and the fit to experimental data can be seen in Figure 4 (main text) and Supplementary Figures S3 and S4 in Sup_1.pdf. We approximate confidence intervals for the locally identifiable parameters according to Equation 6. Although stochastic, Monte-Carlo approaches can provide more accurate estimates of parameter confidence intervals for non-linear models [e.g. Antoniewicz et al., 2006], application of those methods would have prohibitively high computational cost for models of this parameter dimensionality.

For the final parameter set we found that 52 of the 233 parameters are locally identifiable (Sup_2.xls). Of the identifiable parameters, a large proportion was protein abundances. Other identifiable parameters were many of the PTP-1B reaction rate parameters and other various rate parameters, including the EGF on rate. Interestingly, although in Step 5 we expanded the upper and lower bounds for the parameters whose values were on the boundary, another round of minimization did not change the values of these parameters from the original bound values (e.g. k_{on68} and k_{on34}). This is most likely due to the unidentifiability of these parameters.

Although only 52 of the 233 parameters are locally identifiable, this does not mean that the other unidentifiable parameters have no effect on the observables. It simply means that only other *combinations* of parameters can uniquely affect the observables; that is, the effects of locally unidentifiable parameters on the observables are not unique, but may still be large. Additionally, given the extreme difficulty in searching the parameter space to find the final parameter set (~3 processor-years), it is clear that the values of the unidentifiable parameters have a significant impact on our ability to find the values of the identifiable parameters and the model fit. This impact is due to high-order interactions between parameters.

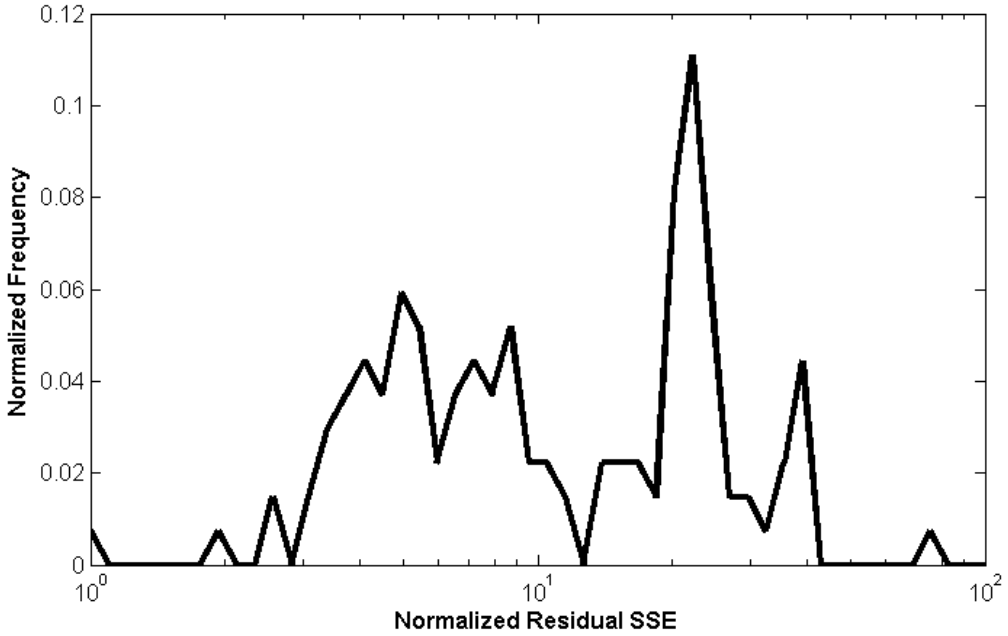

Figure S4-1. Histogram of Candidate Parameter Set Residual Sum of Squared Errors (SSE). Frequency is normalized by the total number of parameter sets (137), and the residual sum of squared errors (SSE) are normalized by the lowest SSE value in the set.

References

Antoniewicz, M. R., Kelleher, J. K. & Stephanopoulos, G. Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. *Metab Eng* **8**, 324-37 (2006).

Polisetty, P. K., Voit, E. O. & Gatzke, E. P. Identification of metabolic system parameters using global optimization methods. *Theor Biol Med Model* **3**, 4 (2006).

Rodriguez-Fernandez, M., Egea, J. A. & Banga, J. R. Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC Bioinformatics* **7**, 483 (2006).

Swameye, I., Muller, T. G., Timmer, J., Sandra, O. & Klingmuller, U. Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. *Proc Natl Acad Sci U S A* **100**, 1028-33 (2003).

Tucker, W., Kutalik, Z. & Moulton, V. Estimating parameters for generalized mass action models using constraint propagation. *Math Biosci* (2006).