APPLICATION OF ROBUST MODEL VALIDATION USING SOSTOOLS TO THE STUDY OF G-PROTEIN SIGNALING IN YEAST

Tau-Mu Yi*
University of California, Irvine
Irvine, CA 92697

Maryam Fazel, Xin Liu, Tosin Otitoju, Jorge Goncalves, Antonis Papachristodoulou,
Stephen Prajna and John Doyle
California Institute of Technology
Pasadena, CA 91125

Abstract

Two major methodological challenges in modeling biological systems are model (in)validation and parameter estimation. The traditional approach is to fit the model parameters to data. An alternative approach pioneered by Packard, Frenklach, Seiler and colleagues (Frenklach et al., 2002) defines the range of parameter values that is consistent with the data while taking into account parametric and data uncertainty. If an invalidation certificate is found, the feasible parameter space is proved empty; otherwise, attempts to describe the feasible parameter space are carried out. We refer to this methodology as Robust Model Validation (RMV). Here we perform RMV using sum of squares (SOS) programs implemented by the MATLAB toolbox SOSTOOLS (Prajna et al., 2002). The principal advantage of SOS over conventional semidefinite programming (SDP) techniques such as the S-procedure is the possibility of using higher-order multipliers to obtain tighter parameter bounds. We applied SOSTOOLS to a simple model of the yeast heterotrimeric G-protein cycle. We were able to invalidate the model based on real experimental data. Furthermore, using synthetic data that did not invalidate the model, we explored different techniques for representing the feasible parameter space.

Keywords

G-protein, signal transduction, biological models, parameter estimation, semidefinite programming.

Introduction

A common complaint among experimentalists is that large mathematical models of biological systems are almost impossible to falsify because they can be made to fit any data by tweaking the many parameters. Even when this practice fails to fit adequately the model to the data, theoretical invalidation is not achieved since it is impossible to simulate the system for all parameter combinations. Furthermore, even if the model is indeed valid, it is hard to estimate parameters when there are inherent uncertainties involved in the model, experiment data, and parameters.

A model invalidation method using surrogate models has been advocated by Packard, Frenklach, Seiler and colleagues (Frenklach et al., 2002). We term this approach Robust Model Validation (RMV). The key idea is to focus on the structure of the model by defining the feasible parameter space instead of specifying a single best (e.g., maximum likelihood) vector of parameter values. Convex relaxations (such as the S-procedure) were used to obtain outer bounds on predictions. The model is invalidated if the feasible parameter space is empty. Whereas Packard and colleagues were more concerned with prediction (Frenklach et al., 2004), we are more interested in model

[*] To whom all correspondence should be addressed

(in)validation and representing the feasible parameter space and parameter relationships. The use of sum of squares (SOS) methodology offers the potential of enhanced accuracy because higher-order relaxations in theory will eliminate the relaxation gap, as well as flexibility in handling any polynomial constraint. The SOS programs were solved by SOSTOOLS (Prajna et al., 2002) in MATLAB.

General Framework and Problem Formulation

Let $u \in \mathbf{R}$, $y \in \mathbf{R}$ denote the input and output of a biological process, and $p \in \mathbf{R}^n$ be the vector of parameters. A model M of the process captures the relation between input and output as $y = M(u,p)$. An experiment provides a dataset (v,d,δ,ε) , where v and d are input and output measurements, and δ and ε are their corresponding uncertainties, such that $|v-u| \le \delta$ and $|d-y| \le \varepsilon$. Let us also assume that prior knowledge on parameters can be expressed by l polynomials. Thus, the prior parameter set is $H = \left\{ p \in \mathbf{R}^n : g_i(p) \le 0, i = 1,\cdots,l \right\}$. The set of parameters that are consistent with m experiments is

$$P = \left\{ p \in H : \left| M(u_i,p) - d_i \right| \le \varepsilon_i, \left| u_i - v_i \right| \le \delta_i, i = 1,\cdots,m \right\} .$$

The problem we consider is to either prove P is empty or describe P as precisely as we can. If P is empty, we can search for a rigorous emptiness proof using barrier certificates (Prajna, 2003). We discuss this approach in a later section. If not empty, P can be described by solutions to a series of optimization problems:

$$\begin{aligned} \Phi^* = \ & \min \quad \Phi(p) \\ & s.t. \quad p \in P \end{aligned} \qquad (1)$$

We choose $\Phi(p)$ depending on the aspect of P we are interested in. For example, $\Phi(p) = p_j$ or $\Phi(p) = -p_j$ for the minimum or maximum values of the j-th parameter.

Unfortunately, the model $y = M(u,p)$ often comes from numerical integration of systems of ODEs like

$$\begin{aligned} \dot{x} &= f(x,u,p), \ x(0) = x_0 \\ y &= g(x,u,p) \end{aligned}$$

where $x \in \mathbf{R}^{n_x}$ are the state variables (e.g., concentration of species), and generally does not have a closed-form expression so that solving (1) is very hard. Therefore, two approximation steps are taken. First, we approximate $M(u,p)$ by a polynomial $S(u,p)$ in the region of interest. $S(u,p)$ is called the surrogate model. The difference between M and S can be made arbitrarily small if the order of S is high enough. We use response surface methods to generate the surrogate model as described in (Frenklach et al., 2002). We assume the approximation error is bounded on the feasible parameter set by e , i.e.

$$\left| M(u_i,p) - S(u_i,p) \right| \le e_i , \quad \forall p \in P, \left| u_i - v_i \right| \le \delta_i, \ \ i = 1,\cdots,m .$$

Then the two sets

$$P_I = \left\{ p \in H : \left| S(u_i,p) - d_i \right| \le \varepsilon_i - e_i, \left| u_i - v_i \right| \le \delta_i, i = 1,\cdots,m \right\} ,$$

$$P_O = \left\{ p \in H : \left| S(u_i,p) - d_i \right| \le \varepsilon_i + e_i, \left| u_i - v_i \right| \le \delta_i, i = 1,\cdots,m \right\} ,$$

can be constructed with $P_I \subseteq P \subseteq P_O$. $P_O = \varnothing$ is a sufficient condition for inconsistency between model and data. This sufficient condition can be checked (with SOS programming) with far lower computational cost than finding barrier certificates. If $P \ne \varnothing$, we have the following bounds:

$$\begin{aligned} \Phi_O^* = \min \ & \Phi(p) \\ s.t. \ & p \in P_O \end{aligned} \le \begin{aligned} \min \ & \Phi(p) \\ s.t. \ & p \in P \end{aligned} \le \begin{aligned} \min \ & \Phi(p) = \Phi_I^* \\ s.t. \ & p \in P_I \end{aligned} .$$

Since the optimization problems with P_I and P_O are generally non-convex, exact values of Φ_O^* and Φ_I^* in general cannot be found. This motivates a second approximation: we compute an outer bound Φ_O on Φ^* using convex relaxations and an inner bound Φ_I on Φ_I^* using standard local optimization techniques, i.e. $\Phi_O \le \Phi_O^* \le \Phi^* \le \Phi_I^* \le \Phi_I$.

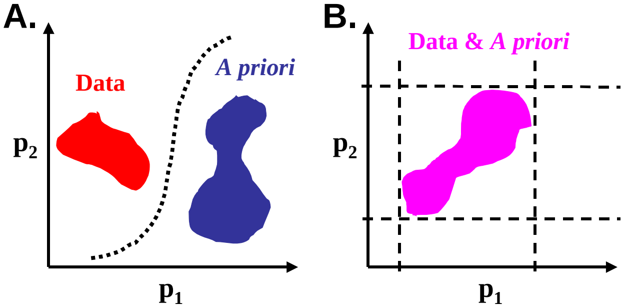

Figure 1. Robust Model Validation (A) A model structure is invalidated if the feasible parameter space (here represented in 2D) is empty. This feasible region can be thought of as the intersection between the parameter region defined by previous knowledge (blue) and the parameter region defined by data constraints (red). A barrier certificate yields a separating function (black dotted line) that proves the intersection is empty. (B) If the feasible parameter space is not empty (magenta), one can attempt to describe this space, for example, by specifying the upper and lower bounds for each parameter. The outer upper and lower bounds are shown (black dashed lines). Other possible representations are described in the text.

Description of Model and Data

We modeled the heterotrimeric G-protein cycle in yeast mediating the response to mating pheromone. The following processes were included in the model: (1) the binding kinetics of ligand (L) to receptor (R); (2) the synthesis and degradation of receptor; (3) activation of G-protein (G) by active receptor (RL); (4) deactivation of Gα-GTP (Ga) catalyzed by the RGS protein Sst2p; and (5) reassociation of the heterotrimer. For several of the reactions, the rate constants or parameters have been measured directly (k_1 , k_2 , k_3 , k_4 , k_5 , G_t). For other reactions, the rate constants were inferred from input-output data (k_6 and k_7), and for some they were based on estimates in the literature (k_8). In this manner, we produced the following ODE model in which the input is the pheromone ligand and the output is the (normalized) level of free $G_{\beta\gamma}$ (Gbg).

$$\begin{aligned} &\text{2a)} \ \ \dot{x}_1 = -k_1 x_1 u + k_2 x_2 - k_3 x_1 + k_5 \\ &\text{2b)} \ \ \dot{x}_2 = k_1 x_1 u - k_2 x_2 - k_4 x_2 \\ &\text{2c)} \ \ \dot{x}_3 = -k_6 x_2 x_3 + k_8 (G_t - x_3 - x_4)(G_t - x_3) \\ &\text{2d)} \ \ \dot{x}_4 = k_6 x_2 x_3 - k_7 x_4 \\ &\text{2e)} \ \ y = (G_t - x_3)/G_t \end{aligned} \qquad (2)$$

where $x_1 = [\mathrm{R}]$, $x_2 = [\mathrm{RL}]$, $x_3 = [\mathrm{G}]$, $x_4 = [\mathrm{Ga}]$, $u = [\mathrm{L}]$, $y = [\mathrm{Gbg}]/G_t$, and G_t = total G-protein.

We have measured the *in vivo* dynamics and regulation of this cycle in yeast using fluorescence resonance energy transfer (FRET) (Yi et al., 2003). Below, we use data from dose-response experiments.

Parameter	u (nM)	y	t (s)
$k_1^0 = 1\mathrm{e}6$	1	0.083	60
$k_2^0 = 1\mathrm{e}-2$	2	0.122	60
$k_3^0 = 4\mathrm{e}-4$	5	0.240	60
$k_4^0 = 4\mathrm{e}-3$	10	0.352	60
$k_5^0 = 4$	20	0.384	60
$k_6^0 = 1\mathrm{e}-5$	50	0.397	60
$k_7^0 = 0.1$	100	0.400	60
$k_8^0 = 1$	1000	0.397	60
$G_t^0 = 1\mathrm{e}4$			

Table 1. Nominal parameter values and experimental input-output data from dose-response experiments.

Results

Direct invalidation of model using barrier certificates

In this section, we examine the problem of invalidating the ODE model (including parameter and output uncertainties) with experimental data, using *barrier certificates*. This method was introduced in (Prajna, 2003), and has been applied to the *E. coli* heat shock response in (El-Samad et al., 2003). It finds functions of state-parameter-time, called barrier certificates, which prove that a model and a feasible parameter set are inconsistent with some time-domain experimental data. Projection of the barrier certificate on the parameter space gives the curve separating the two sets in Figure 1. The search for a barrier can be cast as an SOS program (Prajna, 2003).

We applied this method to the ODE given above along with the first measurement in Table 1. We allowed for 4% uncertainty in k_6 and k_7, and 5% uncertainty in the measured output y=0.083, while fixing all other parameters and the input. Parameters were set to their nominal values except for k_8 (whose nominal value is not known precisely), which was set to 10^{-3}, resulting in a much less stiff system.

Other *a priori* constraints on the states, typically coming from experiments, can also be handled. Here we used simulation to estimate bounds on the states, and included those constraints. We were able to find a barrier certificate *B(x,t,p)*, a polynomial of degree 2 in *x*, 2 in *t*, and 1 in *p*, that invalidates the model with this experiment.

In this approach, there is no need to calculate surrogate models and deal with the extra error introduced by them, making this method more rigorous and less conservative. Also, uncertainty in the initial and final states and state constraints can be explicitly handled. The main disadvantage of the approach is the computational cost. Even though the size of the SOS program (for a fixed-degree barrier) is polynomial in the number of parameters, it grows very fast with the number of input/output data points. Currently, the only known way to include the data requires replicating the states for each data point, adding n_x states to the state space).

Invalidation with surrogate models using SOS

In this section, we use quadratic surrogate models described before to prove that the outer set P_O is empty. This implies the emptiness of P , hence invalidating the original model M. The uncertainty in data is captured by ε and δ . We also include uncertainty in parameters in the form of relative uncertainty τ such that $| k_i - k_i^0 | \le \tau_i \, k_i^0$. In this case, the set P_O is described by polynomials as follows.

$$P_O = \left\{ \begin{array}{l} p = \begin{pmatrix} k_1 & \cdots & k_n \end{pmatrix} \in \mathbf{R}^n : K_j(p) \le 0, V_i(u_i, p) \le 0, \\ \quad W_i(u_i, p) \le 0, Z_i(u_i) \le 0, \; i = 1, \cdots, m, \; j = 1, \cdots, n \end{array} \right\},$$

where $K_j(p) = (k_j - k_j^0 - \tau_j k_j^0)(k_j - k_j^0 + \tau_j k_j^0)$, $V_i(u_i, p) = S(u_i, p) - d_i - \varepsilon_i - e_i$, $W_i(u_i, p) = d_i - \varepsilon_i - e_i - S(u_i, p)$, and $Z_i(u_i) = (u_i - v_i - \delta_i)(u_i - v_i + \delta_i)$. Therefore, if we can find nonnegative multipliers λ 's such that

$$\sum_{j=1}^{n} \lambda_{Kj} K_j(p) + \sum_{i=1}^{m} \left(\lambda_{Vi} V_i(u_i, p) + \lambda_{Wi} W_i(u_i, p) + \lambda_{Zi} Z_i(u_i) \right) > 0$$

for all values of p and u_i , then P_O is proven empty. These λ 's constitute an invalidation certificate. Higher order polynomial multipliers can be found using SOS programs when lower degree polynomials fail to invalidate the model. The values of these multipliers, λ_j for the *j*-th constraint, provide useful information: constraints that have zero multipliers do not contribute to invalidating the model.

In our system, when $\varepsilon_i = 0.1 d_i$, $\delta_i = 0.1 v_i$ for all experiments, $\tau_j = 0.5$ for $j = 6,7$, and $\tau_j = 0.04$ for other parameters, model (2) can be invalidated by the 8 data points in Table 1. We were able to find an invalidation certificate that has positive multipliers corresponding to only 3 data points (1, 7 and 8).

The multipliers corresponding to measurements 1 and 8 were much bigger than that for measurement 7. Therefore, we attempted to invalidate the model using only data points 1 and 8. Model (2) was invalidated by only these two pieces of data. The nonzero multipliers on measurements correspond to constraints $V_1(u_1, p)$ and $W_8(u_8, p)$ -- the upper bound on d_1 and lower bound on d_8 . Hence, directional information can also be obtained: the output measurement in the first experiment was too high and the eighth too low.

Upper and lower bounds

It is a significant challenge to describe the multi-dimensional feasible parameter space. The simplest representation is to calculate the upper and lower bounds on each parameter, thus projecting the space on to the parameter axes (Frenklach et al., 2002).

We created a scenario in which we assumed that the 9 parameters are within prior measured bounds and the data used are synthetic data (Table 2) generated by running simulations of (2) with nominal parameter values. We then used the method proposed in Frenklach et al (2002) to bound feasible parameter set P using hypercubes.

	Outer bounds (measured)	**Input (synthetic)**	**Output (synthetic)**	**Time**
k_1	0.8e6 1.2e6	1	0.0337	60
k_2	0.8e-2 1.2e-2	2	0.0641	60
k_3	3.2e-4 4.8e-4	5	0.1386	60
k_4	3.2e-3 4.8e-3	10	0.2247	60
k_5	3.8 4.2	20	0.3207	60
k_6	7e-6 13e-6	50	0.4116	60
k_7	0.07 0.13	100	0.4373	60
k_8	0.8 1.2			
G_t	0.8e4 1.2e4			

Table 2. Prior measured bounds on parameters and synthetic input-output data.

Using the prior bounds on parameters shown in Table 2, we observed that the outer and inner bounds defined by the data do not further constrain the values of the parameters. Thus, the synthetic data do not add information to prior knowledge on ranges of individual parameters.

Since there is more uncertainty on the values of k_6 and k_7, we expect the synthetic data to have a noticeable impact on feasible values of k_6 and k_7 when all other parameters are fixed to their nominal values. The prior measured uncertainty on k_6 and k_7 are listed in Table 3, as well as outer and inner bounds computed from data. Again the data does not provide additional information on the consistent parameter set P. These results highlight the limitations of bounding each parameter individually, and the importance of capturing parameter correlations.

	Outer bounds (*a priori*)	**Inner bounds (fit to data)**	**Outer bounds (fit to data)**
k_6	5e-6 15e-6	5.13e-6 15e-6	5e-6 15e-6
k_7	0.05 0.15	0.05 0.15	0.05 0.15

Table 3. Comparison (toy scenario) of *a priori* bounds and bounds determined by fitting to input-output data when parameters other than k_6 and k_7 are fixed.

Characterizing parameter interactions

If two parameters belong to different subsystems that do not interact, then the two parameters are expected to be independent of one another. In a 2-d parameter plot, we would see that the feasible parameter region of p_1 would not depend on p_2, and *vice versa* (i.e. the feasible region would be the intersection of horizontal and vertical strips).

More interesting is when the parameters contribute to the same process. For example, the ratio between the G-protein activation rate (k_6) and the deactivation rate (k_7) determines the level of active G-protein, which is measured experimentally. One might expect the ratio k_6/k_7 to tend toward a constant.

Using the synthetic data in Table 2 and the measured bounds in Table 3, we identified two parallel lines (diagonal strip) that bound the 2-d consistent set as tightly as possible from the outside. The slopes and offsets of the lines were set as free variables, and the vertical distance between the two lines as the objective to be minimized. We applied SOS methods to this problem (Figure 2). The shaded region is the intersection of prior knowledge and parameters allowed by data. The consistent parameter set P is a subset of this region. It is clear that although the data does not restrict the value of each parameter, it does only allow certain ratios between k_6 and k_7.

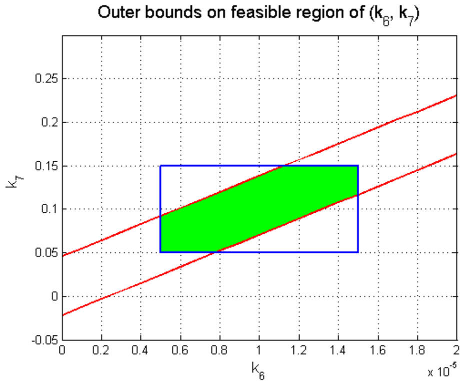

Figure 2. Feasible (k_6, k_7) region allowed by data is inside the shaded region. The prior knowledge (lower and upper bounds) of (k_6, k_7) requires it be inside the rectangle. Synthetic data in Table 2 excludes the region outside the parallel lines.

References

El-Samad, H., Prajna, S., Papachristodoulou, A., Khammash, M. and Doyle, J. C. (2003). Model validation and robust stability analysis of the bacterial heat shock response using SOSTOOLS. In *Proceedings of Conference on Decision and Control*.

Frenklach, M., Packard, A. and Seiler, P. (2002). Predictions from models and data. In *American Control Conference*.

Frenklach, M., Packard, A., Seiler, P. and Feeley, R. (2004). Collaborative data processing in developing predictive models of complex reaction systems. *International Journal of Chemical Kinetics*, **36,** 57-66.

Prajna, S. (2003). Barrier certificates for nonlinear model validation. In *Proceedings of Conference on Decision and Control*.

Prajna, S., Papachristodoulou, A. and Parrilo, P. A. (2002). SOSTOOLS: Sum of squares optimization toolbox for MATLAB. http://www.cds.caltech.edu/sostools/.

Yi, T. M., Kitano, H. and Simon, M. I. (2003). A quantitative characterization of the yeast heterotrimeric G protein cycle. *Proc Natl Acad Sci U S A*, **100,** 10764-9.