Application to insulin signaling-dependent gene regulation model (Sano *et al.*, 2016, *Sci. Signal*)

In this model, insulin regulates gene expression by two-step Michaelis-Menten kinetics: (i) up-regulation of the transcription factor by the insulin signaling, and (ii) up-regulation of the transcription by the transcription factor (Figure 1). I, x, and m represent insulin concentration (ramp stimulation, $I(t) = 0.01 + 0.041625t$ (nM)), insulin signaling-dependent transcriptional regulator's concentration (fold-change), and mRNA concentration of the insulin regulated gene (fold-change), respectively.

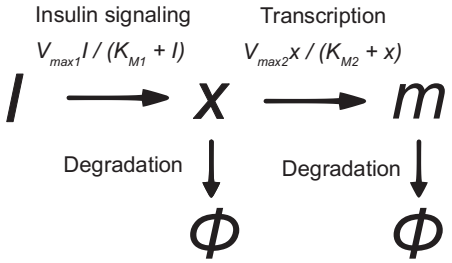


Figure 1. Insulin signaling-dependent gene regulation model.

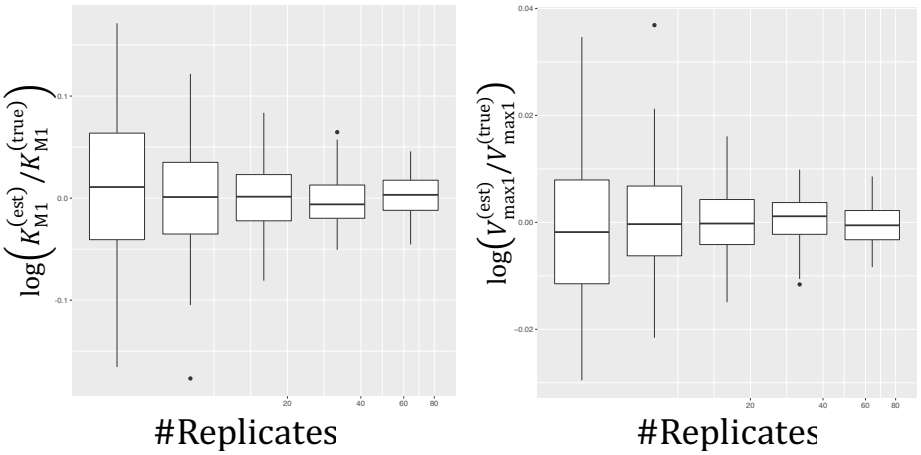
ODE system can be converted to be an identifiable model as follows,

$$\frac{dx}{dt} = V_{\max 1}\left(\frac{I(t)}{K_{\mathrm{M1}} + I(t)} - \frac{I(0)}{K_{\mathrm{M1}} + I(0)}x(t)\right), \quad x(0) = 1$$
$$\frac{dm}{dt} = V_{\max 2}\left(\frac{x(t)}{K_{\mathrm{M2}} + x(t)} - \frac{1}{K_{\mathrm{M2}} + 1}m(t)\right), \quad m(0) = 1.$$

And negative log2 fold change is given as an observable,

$$H(m) = -\log_2 m.$$

First, I evaluated estimation accuracy of our method for this model by twin experiment. Given the true parameters $K_{\mathrm{M1}}^{(\mathrm{true})} = 0.3$ nM, $V_{\max 1}^{(\mathrm{true})} = 1.0$, $K_{\mathrm{M2}}^{(\mathrm{true})} = 5.0$, and $V_{\max 2}^{(\mathrm{true})} = 0.15$, observation data was generated 100 times. For the ideal condition, where observation variance 0.0001 and observation interval 0.5 min, estimation accuracy and confidence intervals were given in Figure 2.



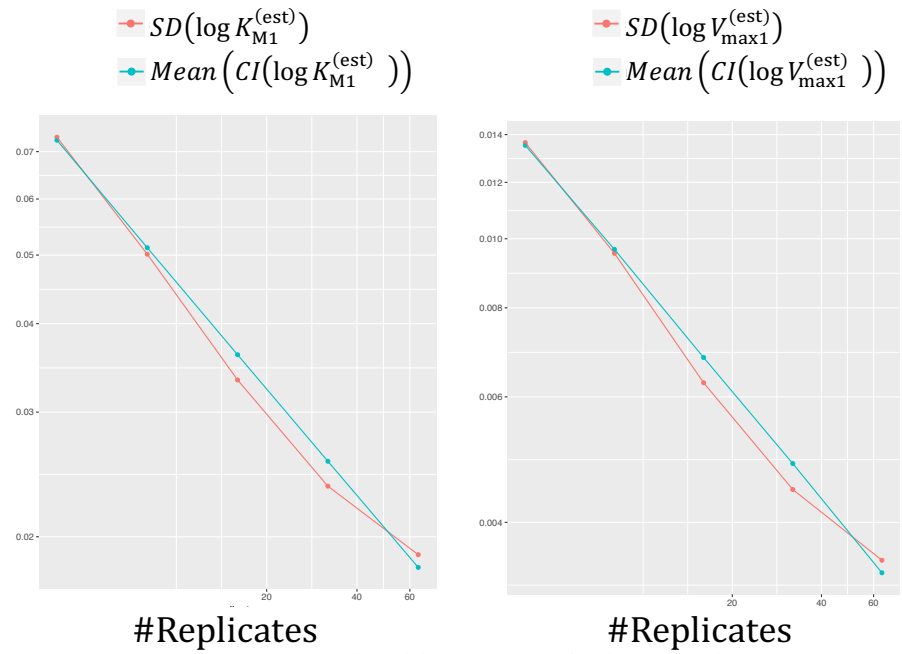
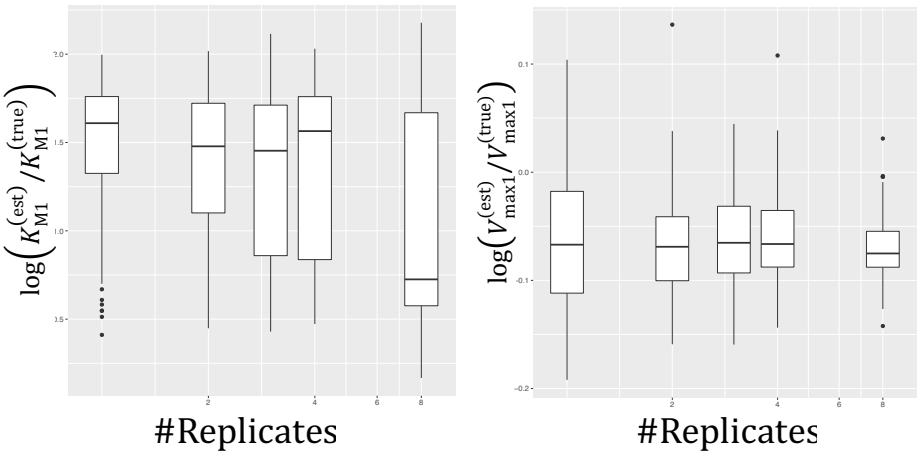


Figure 2. Estimation accuracy of parameters and confidence intervals for the insulin-dependent gene regulation model. observation variance 0.0001, observation interval 0.5 min, #replicates=2-64. SD: standard deviation.

$SD\left(\log K_{\mathrm{M1}}^{(\mathrm{est})}\right)$ can be regarded as an experimentally calculated confidence interval of the estimated K_{M1}, whereas $Mean\left(CI\left(\log K_{\mathrm{M1}}^{(\mathrm{est})}\right)\right)$ was algorithmically computed using second-order adjoint techniques. Thus, correspondence of them assures validity of the confidence interval assessment using our method. We can see that parameter estimation is accurate and confidence intervals are valid in this high-quality data condition. However, for the sparse and rough data condition (observation variance 0.01, observation interval 15 min), which is close to the real data quality, estimation accuracy was low and confidence intervals were not valid (Figure 3).



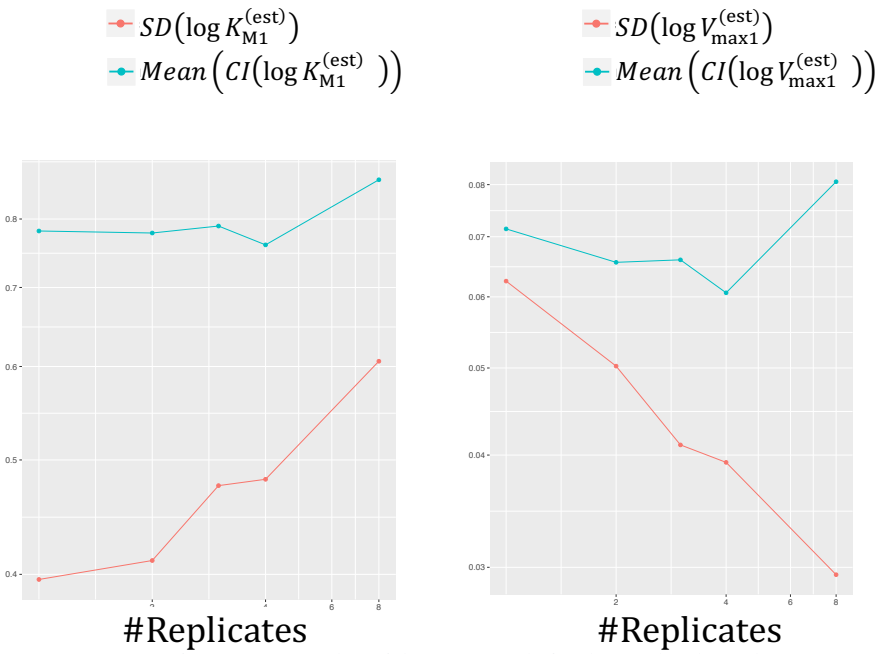


Figure 3. Estimation accuracy of parameters and confidence intervals for the insulin-dependent gene regulation model. observation variance 0.01, observation interval 15 min, #replicates=1-8. SD: standard deviation.

Finally, I applied our method to the real data of *Hmgcr*, which is an insulin up-regulated gene (Sano *et al.*, 2016, *Sci. Signal*). Data assimilation results are given in Figure 4.

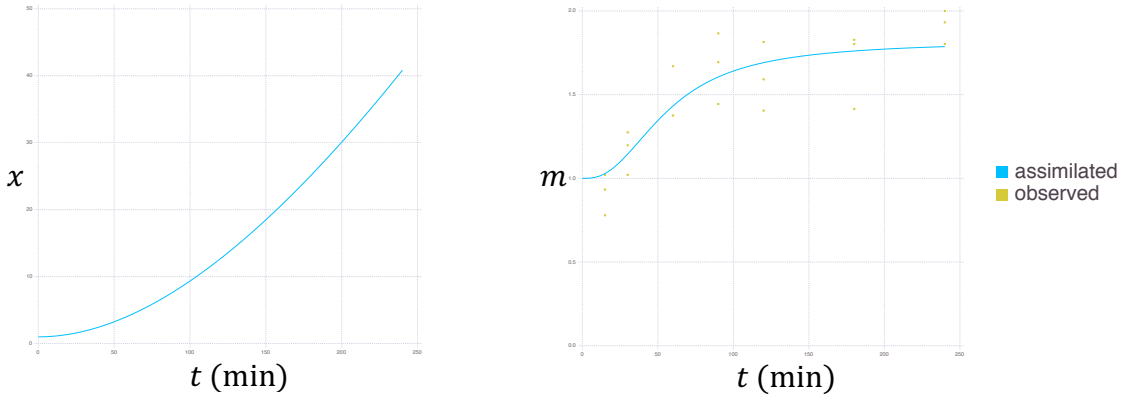


Figure 4. Data assimilation results for Hmgcr.

We can see that the insulin signaling-dependent gene regulation model was successfully assimilated to the experimental observation. However, confidence intervals were large and there was strong positive correlation between $\log K_{\mathrm{M1}}^{(\mathrm{est})}$ and $\log V_{\mathrm{max1}}^{(\mathrm{est})}$, and negative correlation between $\log V_{\mathrm{max1}}^{(\mathrm{est})}$ and $\log K_{\mathrm{M2}}^{(\mathrm{est})}$ (Table1, Figure 5).

Table 1. Estimated parameters and confidence intervals.

	K_{M1}	V_{max1}	K_{M2}	V_{max2}
$p^{(\mathrm{est})}$	4.549	0.190	0.861	0.513
log $p^{(\mathrm{est})}$	1.515	−1.663	−0.149	−0.667
CI (log $p^{(\mathrm{est})}$)	0.579	0.590	0.126	1.258

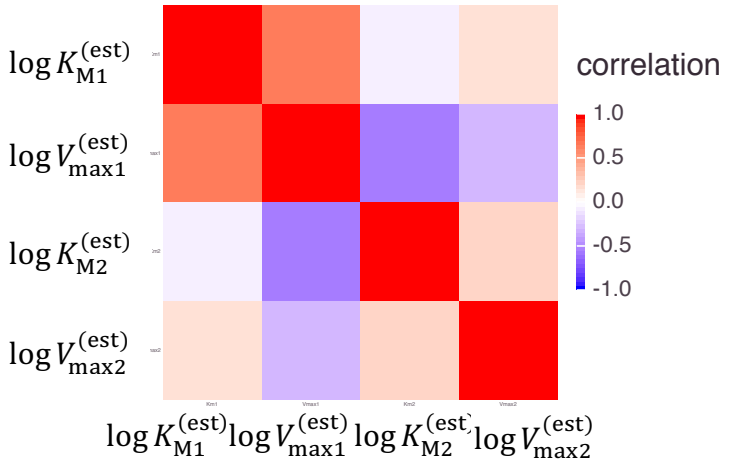


Figure 5. Estimated parameter correlations.

In summary, using synthetic data, our method to estimate ODE parameters and its confidence intervals was validated to be effective as long as rich data was provided. In the *Hmgcr* real data application, the assimilation was successful, however, estimated parameters have low confidence. It seems that better quality of data (more biological replicates and shorter observation interval) are required to determine these parameters with more confidence. Some of the estimated parameters were mutually correlated because some sets of parameters give birth to similar observations. To reduce these correlations and determine each parameter with more confidence, we must manipulate experimental conditions such as input signals and set of observables. Experimental biologists can exploit our method to assess capability for determining parameters under their experimental conditions in advance. Thus, our method can be used to find more effective way to determine ODE parameters.

Application to JAK/STAT signaling model (Maier et al., 2017, *Bioinformatics*)

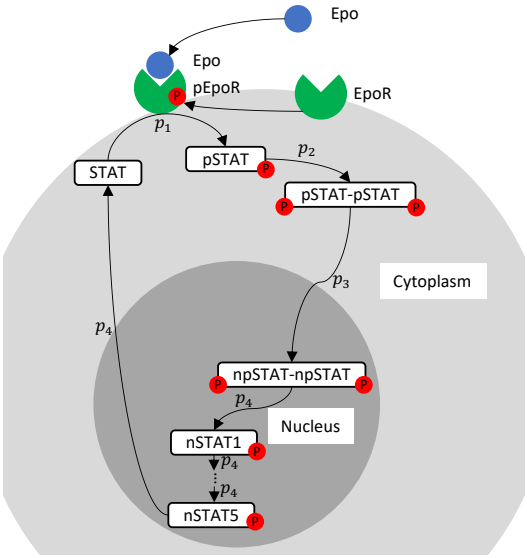


Figure 6. JAK/STAT signaling model.

ODEs and observation model for JAK/STAT signaling model can be expressed as follows, after the parameter conversion for identifiability.

$$\frac{d[\mathrm{STAT}]}{dt} = \frac{\Omega_{\mathrm{nuc}}}{\Omega_{\mathrm{cyt}}} p_4[\mathrm{nSTAT5}] - p_1 u[\mathrm{STAT}]$$
$$\frac{d[\mathrm{pSTAT}]}{dt} = p_1 u[\mathrm{STAT}] - 2p_2[\mathrm{pSTAT}]^2$$
$$\frac{d[\mathrm{pSTAT} - \mathrm{pSTAT}]}{dt} = p_2[\mathrm{pSTAT}]^2 - p_3[\mathrm{pSTAT} - \mathrm{pSTAT}]$$
$$\frac{d[\mathrm{npSTAT} - \mathrm{npSTAT}]}{dt} = \frac{\Omega_{\mathrm{cyt}}}{\Omega_{\mathrm{nuc}}} p_3[\mathrm{pSTAT} - \mathrm{pSTAT}] - p_4[\mathrm{npSTAT} - \mathrm{npSTAT}]$$
$$\frac{d[\mathrm{nSTAT1}]}{dt} = p_4(2[\mathrm{npSTAT} - \mathrm{npSTAT}] - [\mathrm{nSTAT1}])$$
$$\frac{d[\mathrm{nSTAT2}]}{dt} = p_4([\mathrm{nSTAT1}] - [\mathrm{nSTAT2}])$$
$$\frac{d[\mathrm{nSTAT3}]}{dt} = p_4([\mathrm{nSTAT2}] - [\mathrm{nSTAT3}])$$
$$\frac{d[\mathrm{nSTAT4}]}{dt} = p_4([\mathrm{nSTAT3}] - [\mathrm{nSTAT4}])$$
$$\frac{d[\mathrm{nSTAT5}]}{dt} = p_4([\mathrm{nSTAT4}] - [\mathrm{nSTAT5}])$$
$$y_1 = O_{\mathrm{pSTAT}} + [\mathrm{pSTAT}] + 2[\mathrm{pSTAT} - \mathrm{pSTAT}]$$
$$y_2 = O_{\mathrm{tSTAT}} + S_{\mathrm{tSTAT}}([\mathrm{STAT}] + [\mathrm{pSTAT}] + 2[\mathrm{pSTAT} - \mathrm{pSTAT}])$$
$$y_3 = u,$$

where observables y_1 and y_2 stand for concentrations of phosphorylated STAT and that of total STAT in the cytoplasm, respectively. u stands for the concentration of *pEpoR*, which is modeled by cubic spline function $u(t)$ with five parameters $u_1, \cdots, u_5$. Volume ratio of the nucleus to cytoplasm, $\Omega_{\mathrm{nuc}}/\Omega_{\mathrm{cyt}}$ is given as 0.321. $p_1, \cdots, p_4$ are kinetic parameters. $O_{\mathrm{pSTAT}}, O_{\mathrm{tSTAT}}$, and S_{tSTAT} are the offset and scaling parameters, respectively. I evaluated the estimation accuracy of

our method for this model by twin experiment. Data assimilation result and precision matrix for the estimated initial state and parameters is shown in Figure 7.

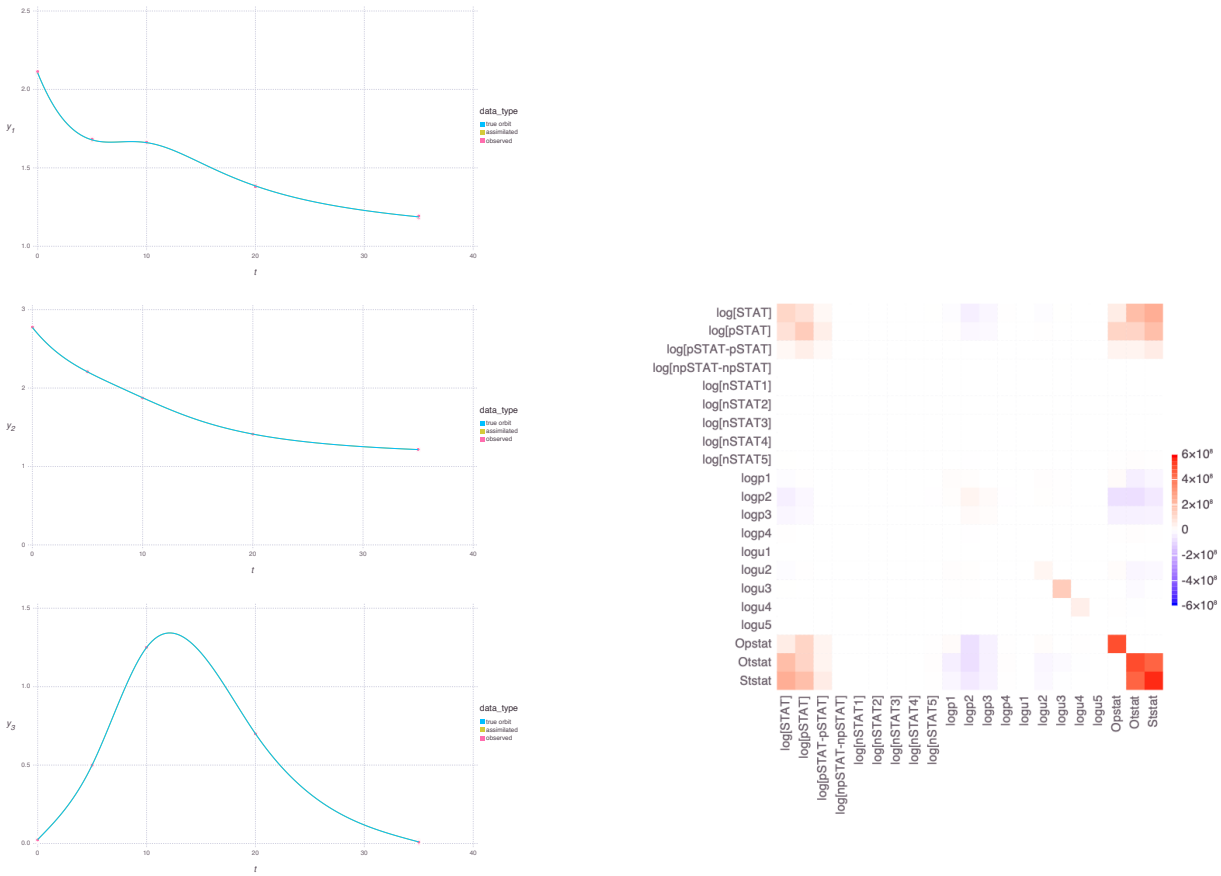


Figure 7. Data assimilation result (left) and precision matrix for estimated initial states and parameters (right) for the JAK/STAT signaling model.

We can see that the model was successfully assimilated to the simulated data. Also, initial state and parameters which directly affect observables, such as [STAT], [pSTAT], [pSTAT − pSTAT], u, O_{pSTAT}, O_{tSTAT}, and S_{tSTAT} had high precision. However, other initial state and parameters which indirectly affect observables had low precision. Estimation accuracy for representative initial state and parameters varying the number of biological replicates is given in Figure 8.

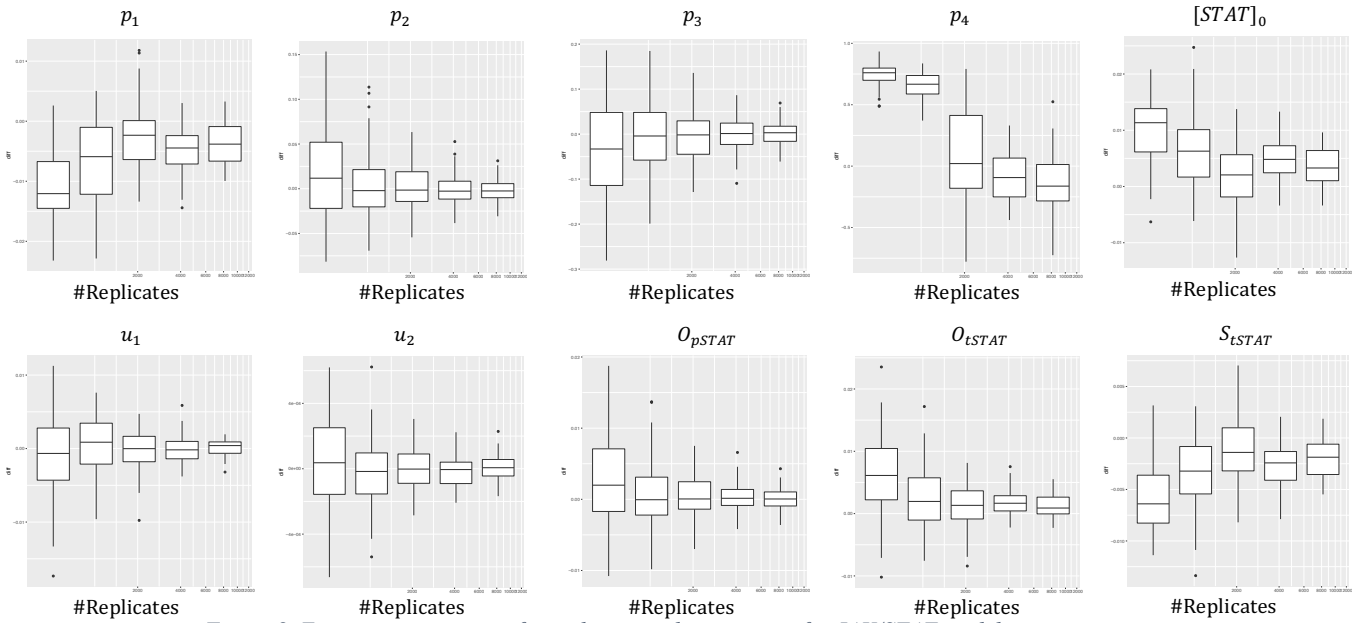


Figure 8. Estimation accuracy of initial state and parameters for JAK/STAT model.

Confidence intervals for representative initial state and parameters varying the number of biological replicates are given in Figure 9.

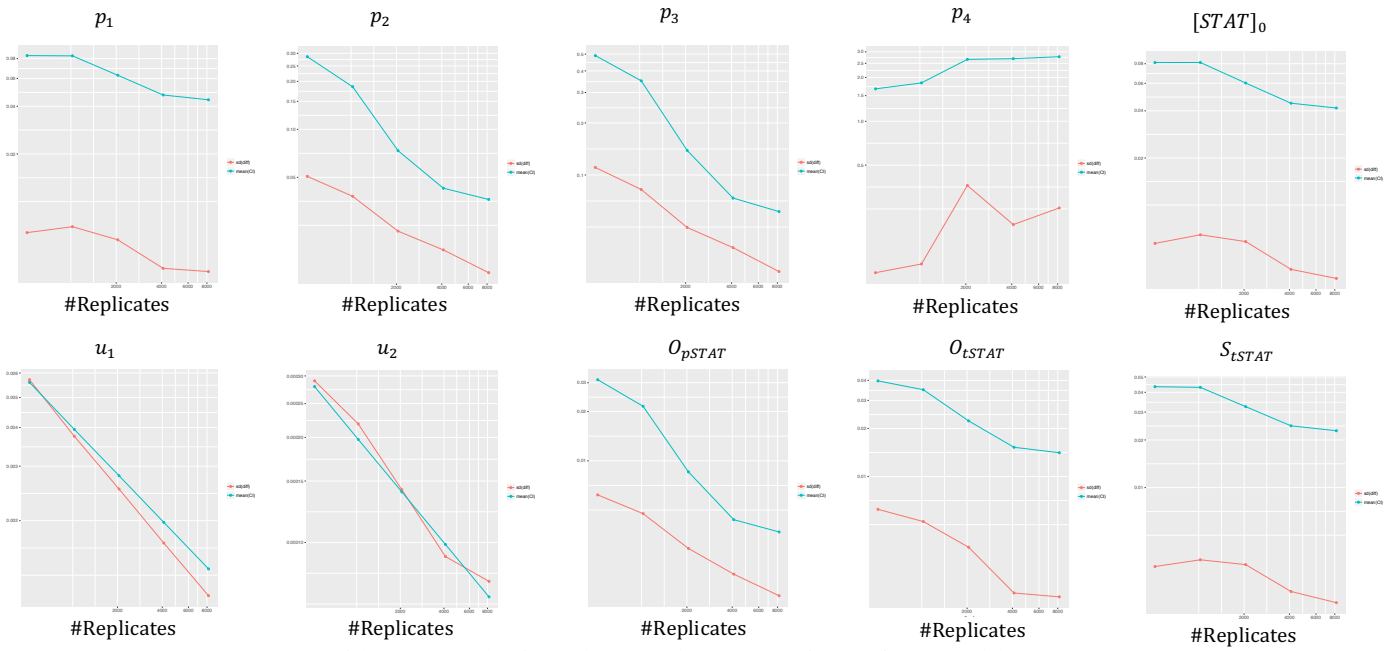


Figure 9. Confidence intervals of initial state and parameters for JAK/STAT model.

We can see that u_1 and u_2, which have direct influence on the observation y_3, were able to be estimated with valid confidence intervals. However, confidence intervals for other parameters could not be properly assessed, possibly due to data shortage and too low identifiability for some model parameters.