Reparametrizing the Sigmoid Model of Gene Regulation for Bayesian Inference^{*}

Martin Modrák¹[0000-0002-8886-7797]

Institute of Microbiology of the Czech Academy of Sciences, Prague, Czech Republic
martin.modrak@biomed.cas.cz

Abstract. This poster describes a novel reparametrization of a frequently used non-linear ordinary differential equation (ODE) model of gene regulation. We show that in its commonly used form, the model cannot reliably distinguish between both quantitatively and qualitatively different parameter combinations. The proposed reparametrization makes inference over the model stable and amenable to fully Bayesian treatment with state of the art Hamiltonian Monte Carlo methods.

Complete source code and a more detailed explanation of the model is available at <https://github.com/cas-bioinf/genexpi-stan>.

Keywords: Gene Regulation · Gene Network Inference · Bayesian Statistics.

1 Introduction

Transcriptional regulation has historically been treated with a vast range of models from Boolean networks, to full Michaelis-Menten simulation of individual chemical reactions. A middle ground is occupied by linear and non-linear ODE models that approximate Michaelis-Menten kinetics. In our attempts to reimplement a Bayesian version of a popular sigmoid ODE model [1] we noticed that very different parameter values - including different sign of the regulatory effect - may result in almost identical behavior, which introduces computational difficulties and questions the interpretation of previous results achieved with the model. Here we describe novel reparametrization of the model implemented in the Stan probabilistic programming language [2] that mitigates these problems.

2 The Model

The ODE model we use is based on [3]. For a bacterial regulon with a single regulator y and targets $x_1 \dots x_n$ the model takes form

$$\frac{dx_i}{dt} = s_i \frac{1}{1 + e^{-\rho_i}} - d_i x_i; \rho_i = w_i y + b_i \quad (1)$$

^{*} Supported by C4Sys research infrastructure (MEYS project No: LM20150055).

Where s_i , w_i , b_i and d_i are parameters to be fit. While bacterial regulons are our primary interest, the model also works with multiple regulators. The model assumes that both the regulator and targets are measured at multiple time points, letting us to solve the ODE. When using microarray data we assume normal observation noise where the standard deviation has two components: one constant and one proportional to the expression. The regulator is fit using B-splines where the spline coefficients are also treated as parameters of the model which let us to both handle uncertainty in regulator measurements and obtain the regulator at arbitrarily small time resolution required for solving the ODE numerically. We note that the model is only weakly identifiable in its direct form (1), as very different parameter sets can result in similar solutions for x_i . Weak identifiability poses computational problems for Bayesian inference, reduces stability of maximum-likelihood estimates and limits interpretability of the model results. Critically, the sign of w_i (whether the regulator is an activator or a repressor) is not well determined for certain target profiles (Fig. 1a) and it may be impossible to determine whether $w_i \simeq 0$ (Fig. 1b). Negligible effect on x_i can also be observed under linear transformations of (w_i, b_i) when $|w_i y + b_i| \gg 0$ (Fig. 1c) and under approximately linear transformations of (s_i, d_i) (Fig. 1d).

Fig. 1. Simulated examples of weakly identified parameters in the sigmoid model. Expression of the regulator (red, solid line) and multiple significantly different parameter values that give raise to similar solutions of the ODE (dashed lines) are shown together with possible measured target profiles (dots) that are insufficient to distinguish between the solutions. **a)** $(w, b, s, d) \in \{(5, -1, 3, 2), (-5, 10, 2.5, 1.4)\}$ **b)** $(w, b, s, d) \in \{5, -1, 3, 3\}, (0, 0, 6, 3.4), (0, 0, 16, 10)$ **c)** $s = 1, d = \frac{1}{2}$ and $(w, b) \in \{(5, -3), (10, -6), (50, -30)\}$ **d)** $w = 1, b = -1$ and $(s, d) \in \{(10, 5), (19, 10), (180, 100)\}$

The aforementioned identifiability problems can be mitigated by a) fixing $I_i = \text{sgn}(w_i)$ and running separate fits when both signs have to be tested and b) replacing w_i, b_i and s_i with μ_i^ρ , the mean regulatory input σ_i^ρ , the std. deviation of the regulatory input and a_i - normalized ratio of s_i to d_i . The original parameters can then be recovered as:

$$w_i = I_i \frac{\sigma_i^\rho}{\text{sd}(y)} \quad (2)$$

$$b_i = \mu_i^\rho - w_i E(y) \quad (3)$$

$$s_i = \frac{a_i d_i \max \tilde{x}_i}{E\left(\frac{1}{1+e^{-\rho_i}}\right)} \quad (4)$$

where E and sd correspond to the sample mean and standard deviation and $\tilde{x}_i$ is the observed expression of gene i .

To our knowledge, this is the first time the identifiability issues are reported and handled for the sigmoid ODE model. Caution should therefore be exercised when interpreting the parameter values found in previous works using this model.

3 Workflow and Remarks

A model fit is in itself insufficient to determine whether a regulation is plausible. To assess plausibility, we test whether the model improves over two baselines: a) constant synthesis ($w_i = 0$) and b) a model where the regulator spline is not constrained by regulator measurements data. We use the LOO-IC criterion [4] to compare the models.

When there are known regulations, the model can be trained by fitting those known regulations first. The posterior estimates of the expression of the regulator can then be used to decrease uncertainty when fitting putative novel targets. Performance-wise, the training phase takes several minutes and fitting 881 putative targets individually on an 8-core machine takes 1 hour which promises faster performance than [1].

References

1. Titsias, M. et al.: Identifying targets of multiple co-regulating transcription factors from expression time-series by Bayesian model comparison. *BMC Systems Biology* 6:53 (2012).
2. Carpenter, B. et al.: Stan: A probabilistic programming language. *Journal of Statistical Software* 76(1) (2017).
3. Vohradsk, J.: Neural Model of the Genetic Network. *Journal of Biological Chemistry* 276(39), 36168-36173 (2001).
4. Vehtari, A. et al.: loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models, <https://cran.r-project.org/package=loo>, (2018).