Parameter Inference in Differential Equation Models of Biopathways using Time Warped Gradient Matching

Mu Niu[(1)], Simon Rogers[(2)], Maurizio Filippone[(3)],Dirk Husmeier[(1)]

(1)School of Mathematics and Statistics, University of Glasgow, UK

(2) Department of Computing Science, University of Glasgow,UK

(3) Department of Data Science, Eurecom, France

Keywords: Biopathways, differential equations, gradient matching, reporoducing kernel Hilbert space, time warping, objective function, optimization.

Abstract. Parameter inference in mechanistic models of biopathways based on systems of coupled differential equations is a topical yet computationally challenging problem, due to the fact that each parameter adaptation involves a numerical integration of the differential equations. Techniques based on gradient matching, which aim to minimize the discrepancy between the slope of a data interpolant and the derivatives predicted from the differential equations, offer a computationally appealing shortcut to the inference problem. However, gradient matching critically hinges on the smoothing scheme for function interpolation, with spurious wiggles in the interpolant having a dramatic effect on the subsequent inference. The present article demonstrates that a time warping approach aiming to homogenize intrinsic functional length scales can lead to a significant improvement in parameter estimation accuracy. We demonstrate the effectiveness of this scheme on noisy data from a dynamical system with periodic limit cycle and a biopathway.

1 Scientific Background

The elucidation of the structure and dynamics of biopathways is a central objective of systems biology. A standard approach is to view a biopathway as a network of biochemical reactions, which is modelled as a system of ordinary differential equations (ODEs). This system can typically be expressed as:

$$\dot{\boldsymbol{x}} = \frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}\left(\boldsymbol{x}(t), \boldsymbol{\theta}\right), \tag{1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_r)$ is a time-dependent vector of r state variables, and the parameters $\boldsymbol{\theta}$ determine the kinetics of the interactions. For complex biopathways, only a small fraction of $\boldsymbol{\theta}$ can typically be measured. Hence, the elucidation of the biopathway dynamics requires the major proportion of kinetic parameters to be inferred from observed (typically noisy and sparse) time course concentration profiles. In principle, this can be accomplished with standard techniques from machine learning and statistical inference. These techniques are based on first quantifying the difference between predicted and measured time course profiles by some appropriate metric, to obtain the likelihood of the data. The parameters are then optimized to maximize the likelihood (or a regularised version thereof). However, the nature of the ODE-based model (1) renders the inference problem computationally challenging in two respects. Firstly, for nonlinear functions $f(.)$, the ODE system (1) usually does not permit closed-form solution. One therefore has to resort to numerical integration every time the kinetic parameters $\boldsymbol{\theta}$ are adapted, which is computationally onerous. Secondly, the likelihood function in the space of parameters $\boldsymbol{\theta}$ is typically *not* unimodal, but suffers from multiple local optima. Hence,

even if a closed-form solution of the ODEs existed, inference by maximum likelihood would be NP-hard, calling for a computationally expensive iterative optimisation .

To circumvent the excessive computational complexity of explicitly solving the ODE system, as described above, various authors have adopted an approach based on gradient matching [Ramsay et al., 2007, Xun et al., 2013, Calderhead et al., 2009, Dondelinger et al., 2013, Macdonald et al., 2015, González et al., 2013, 2014]. The idea is based on the following two-step procedure. In a first *smoothing* step, obtain an estimator of the solution directly from the data. In a second *inference* step, estimate the kinetic parameters $\boldsymbol{\theta}$ by optimizing a functional criteria constructed from the difference between the slope from the estimated solution and the $\boldsymbol{\theta}$-dependent time derivative from the ODEs. In this way, the ODEs never have to be solved explicitly, and the initial conditions do not have to be inferred. A problem intrinsic to this approach is the critical dependence of the inference scheme on the form of the interpolant. Small "wiggles", which are hardly discernible at the level of the interpolant itself, can have dramatic effects at the level of the derivatives, which determine the parameter estimation. For noisy data, an adequate smoothing scheme is essential. Any smoothing scheme is based on intrinsic functional length scales, though, and these length scales may vary in time.

In the present paper, we present a new method that aims to homogenize the intrinsic length scales. The basic idea is that a regular sinusoid is easy to learn, whereas a quasi-periodic signal with varying frequencies is not. The objective, hence, is to find a warping of the time axis that counteracts the inhomogeneity in the period. This can easily be effected in principle. The characteristic feature of a regular sinusoid is the proportionality of the original function to its second derivative. Hence, we need to find a bijective transformation of time such that some metric quantifying the difference between the original function and a rescaled version of its second derivative is minimized in warped time. The procedure thus reduces to a double minimization problem, with respect to both the parameters of the map and the scaling parameter.

2 Materials and Methods

We assume that we have time series of n noisy observations $\boldsymbol{y}_s = (y_{s1}, \ldots, y_{sn})'$ of the states $\boldsymbol{x}_s = (x_{s1}, \ldots, x_{sn})'$, subject to iid additive Gaussian noise $\boldsymbol{\epsilon}_k \sim N(0, \sigma^2 \boldsymbol{I})$:

$$\boldsymbol{y}_s = \boldsymbol{x}_s + \boldsymbol{\epsilon}_s \tag{2}$$

and the objective of inference is to learn $\boldsymbol{\theta}$ from these noisy measurements. We adopt an approach based on reproducing kernel Hilbert spaces (RKHS), where functions are expressed as a linear combination of kernel functions evaluated at the data points

$$x(t) = \sum_{i=1}^{n} b_i k(t, t_i) \tag{3}$$

with $b_i \in \mathbb{R}$ and t_i is the ith time point. In this framework, the unknown concentrations in eq.(1) for the sth component of the dynamical system at time t (which implies $m = 1$) can be modelled as

$$g_s(t; \boldsymbol{b}_s) \;=\; \sum_{i=1}^{n} b_{si} k(t, t_i) \tag{4}$$

with derivatives

$$\dot{g}_s(t; \boldsymbol{b}_s) \;=\; \sum_{i=1}^{n} b_{si} \frac{\partial k(t, t_i)}{\partial t} \;=\; \sum_{i=1}^{n} b_{si} \dot{k}(t, t_i) \tag{5}$$

$$\ddot{g}_s(t; \boldsymbol{b}_s) \;=\; \sum_{i=1}^{n} b_{si} \frac{\partial^2 k(t, t_i)}{\partial t^2} \;=\; \sum_{i=1}^{n} b_{si} \ddot{k}(t, t_i) \tag{6}$$

The ODE parameter $\boldsymbol{\theta}$ can then be estimated by minimizing the difference between $\dot{\boldsymbol{g}}(t_i)$ and the gradient predicted from the ODEs, $\boldsymbol{f}(\boldsymbol{g}(t_i), \boldsymbol{\theta})$, using the following loss function:

$$L(\boldsymbol{\theta}) \;=\; \sum_{s=1}^{r} \sum_{i=1}^{n} \Big[\dot{g}_s(t_i) - f_s(\boldsymbol{g}(t_i), \boldsymbol{\theta}) \Big]^2 \tag{7}$$

In order to overcome the difficulties that variations in intrinsic functional length scales impose on smooth function interpolation, we introduce a two-layer approach. The objective of the first layer is to transform, for each of the variables s of the dynamical system, time t via a bijection $\tilde{t} = w_s(t)$ such that in warped time $\tilde{t}$, the unknown solutions x_s of the dynamical system show less variation in their intrinsic length scales. More specifically, we target oscillating functions and aim to transform them into a regular sinusoid by exploiting the fact that a sinusoid is closed under second-order differentiation (subject to a rescaling). We define the transformation of time as

$$\tilde{t} = w_s(t, \boldsymbol{b}^w, l^w) = \sum_{j=1}^{n} \exp\left(b_j^w\right) \mathcal{S}(t - t_j, l^w); \qquad \mathcal{S}(z, l^w) = \frac{1}{1 + \exp(-l^w z)} \tag{8}$$

where the strict monotonicity of $\mathcal{S}(.)$ and the non-negativity of $\exp(.)$ guarantee bijectivity. The number of basis functions n can, in principle, be treated as a model selection problem. In practice, we found that setting n to the actual number of observations gave satisfactory results (as reported in Section 3). In the original time domain, the sth variable of the dynamical system, $x_s(t)$, is approximated by the smooth interpolant $g_s(t)$. This function is now transformed, by virture of the bijection (8), into $q_s(\tilde{t})$, where

$$g_s(t) = q_s \circ w_s(t) = q_s(\tilde{t}) \tag{9}$$

and $w_s(t)$ is shorthand notation for the bijection defined in (8).

Step 1: Initialization We initialize the system with standard kernel ridge regression. This gives us the smooth interpolants $g_s(t)$ in the original time domain t. We then initialize $\tilde{t} = t$ and $g_s(t) = q_s(\tilde{t})$, for each of the variables s of the dynamical system in turn.

Step 2: Time warping. The bijection between the original time domain $t \in [T_0, T_1]$ and the warped domain $\tilde{t} \in [\tilde{T}_0, \tilde{T}_1]$ is obtained by minimising the objective function

$$L_w = \int \left(\ddot{q}_s(\tilde{t}) + [\lambda^w]^2 q_s(\tilde{t}) \right)^2 d\tilde{t} + \lambda_t \left(\left(\tilde{T}_1 - T_1 \right)^2 + \left(\tilde{T}_0 - T_0 \right)^2 \right) \tag{10}$$

The first term is minimized if $q_s(\tilde{t})$ is a regular oscillation (i.e. phase-shifted cosine or sinusoid) with angular frequency λ^w. In practice, we usually have some prior knowledge about typical periods. This can easily be incorporated by restricting the domain of λ^w, e.g. by modelling it as the output of a rescaled sigmoidal function. The second term is a regularization term, weighted by a penalty parameter $\lambda_t > 0$, to discourage degenerate solutions. The practical choice of λ_t is not critical as long as it is sufficiently large. (The practical procedure is to increase λ_t until the results are invariant wrt a further increase.). The integral in (10) is analytically intractable and needs to be solved numerically:

$$L_w = \sum_{i=1}^{n} \left(\ddot{q}_s(\tilde{t}_i) + [\lambda^w]^2 q_s(\tilde{t}_i) \right)^2 + \lambda_t \left(\left(\tilde{T}_1 - T_1 \right)^2 + \left(\tilde{T}_0 - T_0 \right)^2 \right) \tag{11}$$

The parameters λ^w, l^w and $\boldsymbol{b}^w$ are optimized iteratively until some convergence criterion is met.

Step 3: Interpolation. The second layer deals with function interpolation. The original data points $y_s(t_i)$ are mapped to the warped time points, $y(\tilde{t}_i)$. We then apply standard kernel ridge regression with RBF kernel in the warped domain, which gives us the smooth interpolant $q_s(\tilde{t})$, for each of the variables s in the dynamical system in turn:

$$q_s(\tilde{t}; \boldsymbol{b}_s^q) \;=\; \sum_{j=1}^{n} b_{sj}^q k(\tilde{t}, \tilde{t}_j) \tag{12}$$

Note that this interpolation problem is less susceptible to overfitting or oversmoothing, due to the fact that the initrinsic functional length scales (i.e. periods for an oscillating signal) have been homogenized by virtue of the time warping. Unwarping $q_s(\tilde{t})$ back into the original time domain t is straightforward. Since $w_s(t)$ is bijective, we have $g_s(t) = q_s(\tilde{t})$, and

$$\frac{dg_s(t)}{dt} = \frac{dq_s(\tilde{t})}{dt} = \sum_{j=1}^{n} b_{sj}^q \frac{\partial k(\tilde{t}, \tilde{t}_j)}{\partial \tilde{t}} \frac{d\tilde{t}}{dt} = \sum_{j=1}^{n} b_{sj}^q \frac{\partial k(\tilde{t}, \tilde{t}_j)}{\partial \tilde{t}} w_s'(t) \tag{13}$$

Step 4: Gradient matching. We finally estimate the ODE parameters with gradient matching, i.e. by minimizing the following objective function[1] with respect to $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}) \;= \sum_{s=1}^{r} \sum_{i=1}^{n} \left[\dot{g}_s(t_i) - f_s(\boldsymbol{g}(t_i), \boldsymbol{\theta}) \right]^2 \;=\; \sum_{s=1}^{r} \sum_{i=1}^{n} \left[\frac{dq_s(\tilde{t}_i)}{d\tilde{t}_i} \frac{d\tilde{t}_i}{dt_i} - f_s(\boldsymbol{q}(\tilde{t}_i), \boldsymbol{\theta}) \right]^2 \tag{14}$$

3 Results

The objective of our simulation study is to evaluate the performance of the novel two-level time-warping method proposed in Section 2 with the standard RKHS gradient matching method summarized in Section 1. This method is akin to the one proposed in [González et al., 2013, 2014] and hence representative of the current state of the art. We refer to these methods as RKGW (W for warping) and RKG, respectively. For this comparative evaluation, we have generated time series from one well-known dynamical system and a biopathway. We have repeatedly and independently subjected these data to additive iid Gaussian noise, over a range of signal-to-noise ratios (SNR).

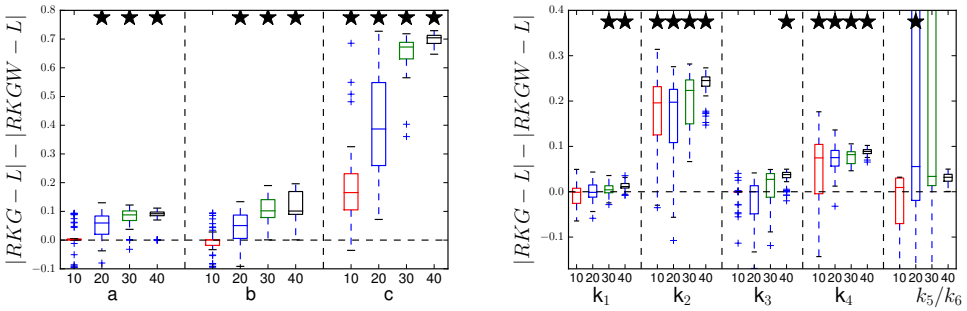
FitzHugh-Nagumo The FitzHugh-Nagumo system is a two-dimensional dynamical system used for modelling spike generation in axons[FitzHugh, 1955]. It has two state variables, x_1 and x_2, and three parameters: a, b and c. We generated equidistant time series of length $n = 37$ from

$$\dot{x_1} = c \cdot \left(x_1 - x_1^3/3 + x_2 \right), \qquad \dot{x_2} = -c^{-1} \left(x_1 - a + b \cdot x_2 \right) \tag{15}$$

Biopathway A model for the interactions of five protein isoforms, S, dS, R, RS, Rpp, in a signal transduction pathway was studied by Vyshemirsky and Girolami [2008], based on mass action and Michaelis-Menten kinetics:

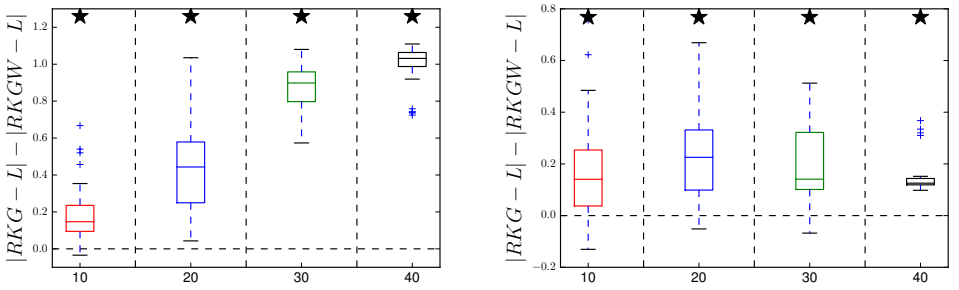
$$\begin{aligned}
[\dot{S}] &= -k_1 \cdot [S] - k_2 \cdot [S] \cdot [R] + k_3 \cdot [RS] \\
[\dot{dS}] &= k_1 \cdot [S] \\
[\dot{R}] &= -k_2 \cdot [S] \cdot [R] + k_3 \cdot [RS] + \frac{k_5 \cdot [Rpp]}{k_6 + [Rpp]} \\
[\dot{RS}] &= k_2 \cdot [S] \cdot [R] - k_3 \cdot [RS] - k_4 \cdot [RS] \\
[\dot{Rpp}] &= k_4 \cdot [RS] - \frac{k_5 \cdot [Rpp]}{k_6 + [Rpp]}
\end{aligned} \tag{16}$$

[1] Recall that t_i depends on s, so a more accurate (but cumbersome) notation would be $g_s(t_i) \to g_s(t_i^s)$.



(a) FitzHugh-Nagumo (b) Biopathway

Figure 1: **Method comparison in parameter space.** The box plots represent, for each true parameter L, the distribution (from 50 independent noise instantiations) of differences between the absolute error of the parameter estimates with the standard method (RKG, Section 1, no warping), and the absolute error of estimates with the proposed method (RKGW, Section 2, with time warping). Positive values (above the dashed horizontal line) indicate that time warping improves performance. The horizontal axis shows different signal-to-noise ratios for each DE parameter. Asterisks above a box indicate where the performance improvement is significant (based on a paired t-test).



(a) FitzHugh-Nagumo (b) Biopathway

Figure 2: **Method comparison in function space.** Similar boxplot representation as in Figure 1, but showing the distribution of the differences between the absolute errors of the function estimates; these function estimates are obtained by inserting the estimated parameters into the ODEs. Positive values indicate that the proposed method outperforms the standard method, asterisks indicate that the improvement is significant (paired t-test).

The results are shown in Figures 1 and 2 and demonstrate that the proposed time warping method achieves a significant improvement in the ODE parameter inference.

4 Conclusion

Carrying out parameter inference in biopathway models described by ODEs is generally difficult due to the need to repeatedly perform computationally expensive numerical integration to solve the ODEs. While gradient matching approaches mitigate this issue, their success critically hinges on the quality of the interpolation scheme. In cases where the solutions to the DE systems exhibit nonstationarity and substantial variations of intrinsic length scales, standard RKHS or Gaussian process approaches typically fail to accurately represent the unkown true functions, leading to poor ODE parameter estimates. In this paper, we have proposed a remedy for this problem by combining gradient matching techniques and time warping. The latter, in particular, is inspired by the work in Calandra et al. [2016], where Gaussian processes are made nonstationary by a reparameterization of the input space. In our work, we use a RKHS interpolation approach instead, and we learn the reparameterization by optimizing a separate objective function that particularly aims to homogenize the intrinsic functional length scales. We have demonstrated that the proposed time warping is effective in improving the quality of

gradient matching approaches in two applications that are representative of biological dynamical systems, one with a limit cycle, the other with a stable equilibrium point.

Our work proposes a first proof of concept that time warping is useful to improve parameter inference in ODE models. We are currently investigating extensions of our work in the direction of including some form of regularization in the estimation of the parameters based on the structure of the ODEs. This could come in the form of alternating the revising of the interpolant in light of the estimated ODE parameters and the estimation of the ODE parameters, or in the form of a prior, following, e.g., the work on hierarchical Bayesian models in Xun et al. [2013].

Acknowledgments

This work was supported by EPSRC (EP/L020319/1).

References

R. Calandra, J. Peters, C. E. Rasmussen, and M. P. Deisenroth. Manifold Gaussian Processes for Regression. *ArXiv e-prints*, February 2016.

Ben Calderhead, Mark Girolami, and Neil D Lawrence. Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 217–224, 2009.

Frank Dondelinger, Maurizio Filippone, Simon Rogers, and Dirk Husmeier. ODE parameter inference using adaptive gradient matching with Gaussian processes. In *Sixteenth International Conference on Artificial Intelligence and Statistics*, 2013.

Richard FitzHugh. Mathematical models of threshold phenomena in the nerve membrane. *The bulletin of mathematical biophysics*, 17(4):257–278, 1955.

Javier González, Ivan Vujačić, and Ernst Wit. Inferring latent gene regulatory network kinetics. *Statistical applications in genetics and molecular biology*, 12(1):109–127, 2013.

Javier González, Ivan Vujačić, and Ernst Wit. Reproducing kernel Hilbert space based estimation of systems of ordinary differential equations. *Pattern Recognition Letters*, 45:26–32, 2014.

Benn Macdonald, Catherine Higham, and Dirk Husmeier. Controversy in mechanistic modelling with Gaussian processes. In *Journal of Machine Learning Research: Workshop and Conference Proceedings*, volume 37, pages 1539–1547. Microtome Publishing, 2015.

Jim O Ramsay, G Hooker, D Campbell, and J Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796, 2007.

Vladislav Vyshemirsky and Mark A Girolami. Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6):833–839, 2008.

Xiaolei Xun, Jiguo Cao, Bani Mallick, Raymond J Carroll, and Arnab Maity. Parameter Estimation of Partial Differential Equation Models. *Journal of the American Statistical Association*, 108(503):37–41, 2013. ISSN 0162-1459. doi: 10.1080/01621459.2013.794730.