DENSITY ESTIMATION OF INITIAL CONDITIONS FOR POPULATIONS OF DYNAMICAL SYSTEMS

Alberto Giovanni Busetto[1], **Bernd Fischer and Joachim Buhmann**

Institute of Computational Science, ETH Zürich, 8092 Zürich, Switzerland

Introduction In the biological sciences, time series can now be routinely collected from experiments. With time-continuous models, formalization can be based on systems of differential equations with stochastic noise sources. Models often cannot be fully understood but additional domain knowledge might be available.

In general, observations are performed at discrete times and are often scarce and always noisy. The presence of outliers requires robustness, while underdetermination requires additional knowledge. Thus, incorporating prior information is not only useful, but also necessary. Moreover, depending on the involved experimental techniques, direct measurements might not be feasible at the single-cell level. Instead, averaged behaviors over large populations can be measured. Due to experimental costs and required time, the number of sample points is very limited and an optimal selection of measurement points is desirable.

An interesting example is Green Fluorescent Protein (GFP) degradation. In fact, a significant behavioral discrepancy between *in vivo* and *in vitro* data has been observed. In the former case data showed a first-order degradation kinetics [1], while in the latter a zeroth-order kinetics [3]. This distortion could be caused by the discrepancy between single-cell and averaged dynamics. Further research suggests that the initial protein distribution plays an important role in the exhibited discrepancy [5].

Purpose A computational approach that estimates the probability density of the initial conditions for a population of dynamical systems is presented. Its scope extends to a family of problems which includes the described protein degradation example. It permits the formulation of hypotheses that can justify the discrepancy between single-cell and population dynamics. The approach is based on a preprocessing regression that permits the incorporation of domain knowledge. This knowledge is given under the form of prior information about the trajectory of a single cell and about the dynamical behavior of the noisy observations. In similar problems, additional knowledge can be available as a prior over functions. This advantage is not possible with purely data-driven approaches and, when existing, it must be exploited. In systems biology, the chemical reactions are often understood quite well, but complex systems or networks are still under investigation. However, integration of prior knowledge comes with an high cost and, in general, feasible approaches to compute inference must be approximated.

The time-continuous system can only be observed at a finite and small number of sampling points. Time sampling is usually performed uniformly or according to experimental convenience. However, it is highly desirable to compute an optimal sampling, such that the information gain is maximized. The result is relevant especially in the common case of data scarcity and it permits the optimization of the experiment design. Finally, the approach leads to the definition of an optimization method, whose aim is the selection of the optimal set of sampling time points. It provides practical information for experimentalists, establishing a loop between experimentation, modeling and computation.

The underlying assumptions are the following: a cell population consists of a large but finite number of independent and deterministic units, which are parametrically homogeneous but exhibit heterogeneous initial conditions. Furthermore, the cardinality of the set of sampling points is assumed to be fixed.

Methods The dynamical evolution of a population of size s is modeled as an initial value problem

$$\frac{dx(t,x_{0i})}{dt} = f(x(t,x_{0i}),t,\boldsymbol{\theta}), \quad \text{with initial conditions} \quad x(t_0,x_{0i}) = x_{0i} \quad i = 1,2,\ldots,s \tag{1}$$

restricted to the interval $[t_0, t_f]$ and with parameters $\boldsymbol{\theta}$. Assume that the conditions of the Picard-Lindelöf theorem are satisfied. Let $p(x_0)$ be a probability density for the i.i.d. single-cell initial conditions x_{0i}. The measured behavior of the population, for $s \to \infty$, is given by $z_\varepsilon(t) = \mathbb{E}_{x_0}[x(t,x_0)] + \varepsilon(t)$, where $x(t,x_0)$ is the solution for a single-cell. The additive noise that corrupts the measurements is denoted by $\varepsilon(t)$.

The prior information about the population dynamics and the observational errors should be approximated by the selection of an appropriate regression model, finally providing $\tilde{z}(t)$. Since the potential existence of outliers cannot be denied, a robust approach must be employed. This goal can be achieved through Bayesian regression with a mixture of regular observations and outliers. For these likelihood models, analytic inference becomes intractable and various approximation techniques have been proposed. Between these, expectation-propagation and Markov Chain Monte Carlo are able to produce satisfactory results [4].

[1]email: busettoa@inf.ethz.ch

Let n be the dimension of a Parzen window estimator $\widehat{p}_n(x_0, \mathbf{p}) \simeq p(x_0)$. With the introduced approximation, one has

$$\tilde{z}(t) \simeq \int \widehat{p}_n(x_0, \mathbf{p}) x(t, x_0) dx_0 = \sum_{i=1}^{n} p_i \int w_i(x_0) \tilde{x}(t, x_0) dx_0 = \sum_{i=1}^{n} p_i \tilde{\phi}_i(t), \tag{2}$$

with a window $w_i(x_0)$ and $\tilde{x}(t, x_0)$ numerically integrated. Then, the average behavior of $w_i(x_0)$ is approximated with a Monte Carlo approach. The integral equation (2) is finally discretized as $\tilde{\mathbf{z}} \simeq \tilde{\mathbf{\Phi}}\mathbf{p}$, where $\tilde{\mathbf{\Phi}} = [\tilde{\phi}_{ji}] = [\tilde{\phi}_i(t_j)] \in \mathbb{R}^{m \times n}$ and $\tilde{\mathbf{z}} = [\tilde{z}_j] = [\tilde{z}(t_j)] \in \mathbb{R}^m$. With a maximum likelihood estimator and considering a mean square error, the problem can be stated as a least squares problem subject to linear inequalities and linear equalities. The formulation is the following: find $\mathbf{p}^*$ such that

$$\mathbf{p}^* = \arg \min_{\mathbf{p}^* \in \mathbb{R}^n_{\geq 0}} \left[\|\tilde{\mathbf{\Phi}}\mathbf{p} - \tilde{\mathbf{z}}\|_2^2 \right] \quad \text{subject to} \quad \sum_{i=1}^{n} p_i = 1 \tag{3}$$

This estimation is sensitive to outliers and, therefore, must be based on the robust regression preprocessing mentioned before.

Now, if $\tilde{\mathbf{\Phi}}$ is full-column rank, then the solution exists and is unique; also, it can be computed in polynomial time with an active set method. Note that, in general, the constraints cannot be relaxed, since the cases where they are active is rather common. Otherwise, when $\tilde{\mathbf{\Phi}}$ is not full-column rank, the solution is not unique. Due to the frequent constraint of data scarcity, this can be often caused by undersampling. In that case, in order to assume the least biased distribution, the information entropy is maximized. The formalization is the following: find $\mathbf{p}^*$ such that

$$\mathbf{p}^* = \arg \max_{\mathbf{p} \in \mathcal{S}} \left[- \int \widehat{p}_n(x_0, \mathbf{p}) \log \left(\widehat{p}_n(x_0, \mathbf{p}) \right) dx_0 \right], \tag{4}$$

where $\mathcal{S}$ is the set of solutions of (3).

Assume that the set of possible sampling time points has a given fixed cardinality l. Then the goal is to find, if possible, a global minimizer such that the information gain is maximized. This can be done by minimizing the differential information entropy of the solution. More formally, the problem is the following: find $\mathbf{t}^*$ such that

$$\mathbf{t}^* = \arg \min_{\mathbf{t} \in [t_0, t_f]^l} \left[- \int \widehat{p}_{n,\mathbf{t}}(x_0, \mathbf{p}) \log \left(\widehat{p}_{n,\mathbf{t}}(x_0, \mathbf{p}) \right) dx_0 \right], \tag{5}$$

where $\widehat{p}_{n,\mathbf{t}}(x_0, \mathbf{p})$ is given by the solution of minimization (3) under the sampling defined by $\mathbf{t}$.

Results and Outlook This approach has been tested on simulated data with known density of initial conditions. During the numerical experimentation, we used the biologically realistic function for molecular decay [2]

$$f(x(t, x_0), t, \gamma, c, \delta, K, V) = \frac{c}{\gamma} \exp(-\gamma t) - \delta x(t, x_0) - \frac{V x(t, x_0)}{K + x(t, x_0)}, \tag{6}$$

which models GFP protein degradation. The results are encouraging, since the ground truth function has been recovered almost perfectly with little prior information. Furthermore, biological data show the importance of this approach to recover the true singe-cell dynamical behavior. In fact, this can be easily masked by densities of initial distributions with long tails. The method can be extended to populations with heterogeneous parameters and inter-cellular time-shifts. In fact, time-shifted behavior can be parametrized and parameters can be formalized as extended initial conditions.

In practice, taking prior information into account is strongly beneficial. Approximate inference permits a feasible approximation of the robust regression, extending the applicability of the whole approach. While the selection of a double model for outliers and regular observations seems promising, which approximation technique provides the best results in this study case is still an open question. This is an issue that deserves attention due to its scientific impact.

Finally, the numerical determination of the optimal subsampling $\mathbf{t}^*$ helps the improvement of the results. It can be combined iteratively in a loop of computation and experimentation, providing further refinements.

References

[1] J. B. Andersen *et al.* New unstable variants of Green Fluorescent Protein for studies of transient gene expression in bacteria. *Appl Environ Microbiol*, 64:2240–2246, 1998.

[2] C. Grilly *et al.* A synthetic gene network for tuning protein degradation in Saccharomyces cerevisiae. *Mol Sys Biol*, 3, 2007.

[3] G. L. Hersch *et al.* SspB delivery of substrates for ClpXP proteolysis probed by the design of improved degradation tags. *PNAS*, 101(33):12136–12141, 2004.

[4] M. Kuss *et al.* Approximate inference for robust Gaussian process regression. Technical Report 136, Tübingen, Germany, 2005.

[5] W. W. Wong *et al.* Single-cell zeroth-order protein degradation enhances the robustness of synthetic oscillator. *Mol Sys Biol*, 3, 2007.