Identifiability analysis and experimental design for dynamical models in systems biology

Andreas Raue, J. Timmer

Angaben zur Veröffentlichung / Publication details:

Raue, Andreas, and J. Timmer. 2010. "Identifiability analysis and experimental design for dynamical models in systems biology." In 6th Workshop on Computation of Biochemical Pathways and Genetic Networks: a BIOMS event; Bioquant, University of Heidelberg, September 16-17, 2010, edited by Katrin Hübner, Tim Johann, Ursula Kummer, and Jennifer Levering, 15–21. Berlin: Logos-Verlag.

Identifiability Analysis and Experimental Design for Dynamical Models in Systems Biology

Andreas Raue[1] and Jens Timmer[1,2]
[1] Physics Institute, University of Freiburg
[2] Freiburg Institute for Advanced Studies, University of Freiburg
79104 Freiburg, Germany

Abstract

Mathematical models of the dynamics of cellular processes promise to yield new insights into the underlying biology and their systems' properties. Since the processes are usually high-dimensional and time-resolved experimental data of the processes are sparse, parameter estimation faces the challenges of structural and practical non-identifiability of the parameters. Non-identifiability induces non-observability. Non-observability reduces the predictive power of the models. We will discuss a new approach that allows for identifiability and observability analysis. We demonstrate how identifiability analysis combined with observability analysis supports the iterative cycle between modelling and experimentation in systems biology.

Introduction

In the context of signaling networks, ordinary differential equation (ODE) systems are frequently used to investigate the dynamic properties of pathway components and their transient modifications. This assumes that diffusion is fast compared to reaction rates and cell volume.

The model equations

$$\dot{\vec{x}}(t,\theta) = \vec{f}(\vec{x}(t,\theta),\vec{u}(t),\theta) \tag{1}$$

$$\vec{y}(t_i,\theta) = \vec{g}(\vec{x}(t_i,\theta),\theta) + \vec{\varepsilon}_i \tag{2}$$

describe via the ODE system (1) the dynamics of n species $\vec{x}$ such as concentrations of proteins in different phosphorylation states. Their dynamical behavior may depend on an input function $\vec{u}(t)$ and model parameters $\theta = \{\theta_1 \dots \theta_l\}$. The species are mapped to m model observables $\vec{y}$, the quantities accessible by experiments, via an observation function $\vec{g}$ in (2). Often, only a subset or combinations of the modeled species are accessible by experiments, meaning that $m < n$. The distribution of the measurement noise $\varepsilon_{ki} \sim \mathrm{N}(0, \sigma_{ki}^2)$ is assumed to be known.

The aim is to match the mathematical model with experimentally observed time-series data, to reconstruct and validate the network structure and to predict model dynamics. An important step is the estimation of beforehand unknown model parameters determining the dynamical behavior. Intrinsically, the outcome of model predictions depends on the estimated model parameters and their *identifiability*. If model parameters θ are *non-identifiable*, meaning that they are not well determined, some parts of the predicted model dynamics are also not specifiable, i.e. some components of $\vec{x}(t,\theta)$ may be *non-observable*. Consequently, the questions that should be answered by the model might not be addressable given the experimental data available. Inferring how identifiability and observability problems can be resolved by additional experimental data is the subject of *experimental design*.

Method

Commonly, many model parameters θ are unknown and have to be estimated from experimental data. The agreement of experimental data $y_k^\dagger(t_i)$ with the observables predicted by the model $y_k(t_i,\theta)$ for parameters θ is measured by an *objective function*, commonly the weighted sum of squared residuals

$$\chi^2(\theta) = \sum_{k=1}^{m} \sum_{i=1}^{d_k} \frac{1}{\sigma_{ki}^2} \left(y_k^\dagger(t_i) - y_k(t_i,\theta) \right)^2 \tag{3}$$

where d_k denotes the number of data-points for each observable $k = 1 \dots m$, measured at time points t_i with $i = 1 \dots d_k$. σ_{ki} are the corresponding measurement errors that are assumed to be known. The parameters can be estimated by finding the parameter values $\hat{\theta}$ that minimize $\chi^2(\theta)$. For normally distributed measurement noise this corresponds to maximum likelihood estimation, see e.g. [1].

The key point is that it is not sufficient to rely on the mere estimated parameter values and their corresponding prediction for the system dynamics. It

is important to consider the uncertainties in the parameter estimation procedure: from measurement uncertainties, to parameter uncertainties and possibly non-identifiabilities, to uncertainties in the predicted model dynamics and possibly non-observabilities. Uncertainties in the parameter estimates are usually described by *confidence intervals*. A confidence interval $[\sigma_i^-, \sigma_i^+]$ of a parameter estimate $\hat{\theta}_i$ to a confidence level $1-\alpha$ signifies that the true value θ_i^* is located within this interval with probability $1-\alpha$.

Identifiability

A parameter θ_i is *structurally identifiable*, if its estimate $\hat{\theta}_i$ is a unique minimum of $\chi^2(\theta)$. It is *practically identifiable*, if the confidence interval of its estimate has finite size. A non-identifiable parameter indicates that it cannot be estimated from the experimental data and hence its confidence intervals are infinite.

An approach for identifiability analysis utilizing the *profile likelihood*

$$\chi^2_{PL}(\theta_i) = \min_{\theta_{j\neq i}} \left[\chi^2(\theta)\right] \tag{4}$$

was proposed by [2]. The idea of the approach is to detect flatness of the likelihood by exploring the parameter space for each parameter in the direction of least increase in $\chi^2(\theta)$. Therefore, for each parameter θ_i individually a section along the minimum of the objective function with respect to all of the other parameters $\theta_{j\neq i}$ is computed. At the same time, the profile likelihood enables to calculate likelihood-based confidence intervals [3, 4].

Structural non-identifiability A structural non-identifiability arises from the model structure only and is independent of the amount and quality of experimental data, see in [5]. Assuming ideal measurements, with arbitrarily many and perfectly chosen measurement time points t_i and absence of measurements errors $\vec{\varepsilon}_i = 0$, the crucial question is whether the model parameters θ are uniquely estimable from the model observables $\vec{y}(t_i, \theta)$.

The analytical solution of $\vec{y}(t_i, \theta)$ may contain an ambiguous parameterization with respect to θ, arising from an insufficient mapping function $\vec{g}$ in (2) that is characterized by functional relations $\vec{h}(\theta_{sub}) = 0$ of a subset of parameters $\theta_{sub} \subset \theta$. In terms of likelihood, a structural non-identifiability manifests as iso-χ^2 manifold

$$\left\{\theta \,|\, \vec{h}(\theta_{sub}) = 0\right\} \quad \Rightarrow \quad \chi^2(\theta) = const. \tag{5}$$

For a two dimensional parameter space, a structural non-identifiability can be visualized by a perfectly flat valley that is infinitely extended along the corresponding functional relation, as illustrated in Fig. 1a by the dashed line. Correspondingly,

this can be detected by a flat line of the profile likelihood for each parameter of θ_{sub}, see Fig. 1b. Consequently, structural non-identifiable parameters are not uniquely identified by measurements of $\vec{y}(t_i, \theta)$ and confidence intervals of $\theta_i \in \theta_{sub}$ are infinite. A parameter is structurally identifiable, if a unique minimum of $\chi^2(\theta)$ with respect to θ_i exists, see Fig. 1c-f.

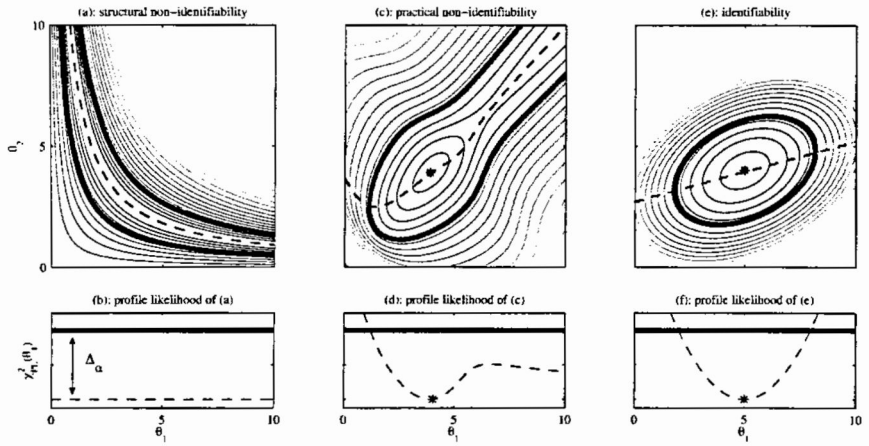

Figure 1: Assessing parameter identifiability of parameter θ_1 from the profile likelihood $\chi^2_{PL}(\theta_1)$. Contour lines in (a,c,e) shaded from black to white correspond to low respectively high values of $\chi^2(\theta)$. Thick lines indicate the threshold Δ_α utilized to asses likelihood-based confidence intervals and asterisk correspond to the optimal parameters $\hat{\theta}$. Dashed lines indicate the profile likelihood for θ_1 and its corresponding trace in (a,c,e).

Practical non-identifiability A parameter that is structurally identifiable may still be practically non-identifiable. This can arise due to insufficient amount and quality of experimental data or inappropriately chosen measurement time points. It manifests in a confidence interval that is infinite, although the likelihood has a unique minimum for this parameter. Confidence intervals can be defined by a threshold Δ_α in the likelihood. This threshold defines a confidence region

$$\{\theta \,|\, \chi^2(\theta) - \chi^2(\hat{\theta}) < \Delta_\alpha\} \quad \text{with} \quad \Delta_\alpha = \mathrm{Q}(\chi^2_{df}, 1-\alpha) \tag{6}$$

whose borders represent *likelihood-based* confidence intervals [6]. The threshold Δ_α is the $1-\alpha$ quantile of the χ^2_{df}-distribution. The choice of df yields confidence intervals that hold jointly for df number of parameters [7], often $df = 1$ is desired.

According to [2], a parameter is practically non-identifiable, if the likelihood-based confidence region (6) is infinitely extended in direction of θ_i indicated by the likelihood staying below a desired Δ_α. Similar to structural non-identifiability, the flattening out of the likelihood can continue along a functional relation. For a two dimensional parameter space, a practical non-identifiability can be visualized as a relatively flat valley, which is infinitely extended, cf. Fig. 1c. This can be detected by the corresponding profile likelihood in Fig. 1d, indicating that the height distance of the valley bottom to the lowest point at $\hat{\theta}$ never exceeds Δ_α. By increasing amount and quality of measured data and/or the choice of measurement time points t_i, a practical non-identifiability will ultimately be remediated, yielding finite confidence intervals, see Fig. 1e-f.

Observability

The uncertainty of parameter estimates $\hat{\theta}$ indicated by non-identifiability directly translates to uncertainty of model trajectories indicated by non-observability. For structurally non-identifiable parameters, the species $\vec{x}$ affected by θ_{sub} can be non-observable, whereas the model observables $\vec{y}$ are by definition invariant. In contrast, for practical non-identifiable parameters, the model observables $\vec{y}$ are affected but stay in agreement with the uncertainties of the experimental data because the likelihood stays below the threshold Δ_α. Nevertheless, some species $\vec{x}$ might be affected strongly by a practical non-identifiability and hence might be not-observable. Also, confidence intervals of parameter estimates translate to confidence intervals of model trajectories.

Experimental Design

Since structural non-identifiability is independent of the accuracy of experimental data, it cannot be resolved by increasing amount and quality of existing measurements. The only remedy is a qualitatively new measurement which alters the mapping function $\vec{g}$ in (2), usually by increasing the number of observed species. For practical non-identifiability, increasing amount and quality of existing measurements may be sufficient but is often not very efficient.

To plan new experiments that efficiently resolve non-identifiability problems, [2] proposed to investigate the set of trajectories along the profile likelihood of θ_i. This corresponds to the observability of the trajectories and reveals spots where the uncertainty of θ_i has the largest impact. Additional measurements at these spots promise to resolve both structural and practical non-identifiabilities and narrow confidence intervals most efficiently. Furthermore, the amplitude of variability of the trajectories at these spots allows to assess the necessary measurement precision to provide adequate data.

Summary

Parameter non-identifiability arises frequently in Systems Biological applications and are often insufficiently considered. We illustrated that parameter identifiability, both structural and practical, and confidence intervals of parameter estimates are a matter of flatness of the likelihood. For structural identifiability, it is critical which and how many of the modeled species can be measured directly. For practical identifiability, also the amount and quality of experimental data and the choice of the measurement time points play an important role. Both causes are inherent to biological applications where experiments are time and cost consuming. Reliable parameter estimates are nevertheless critical before model predictions can be trusted. We discussed a method suitable for systems biological applications that allows to assess and improve both structural and practical parameter identifiability in order to improve the reliability of model predictions.

Acknowledgments

We thank Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling and Ursula Klingmüller for their contributions, and the German Federal Ministry of Education and Research (LungSys 0315415E) and the Excellence Initiative of the German Federal and State Governments (EXC 294) for their funding.

References

[1] Seber, G. & Wild, C., *Nonlinear Regression.* London: J. Wiley, 2003.

[2] Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U. & Timmer, J., Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

[3] Murphy, S. & van der Vaart, A., On profile likelihood. *J. Am. Stat. Assoc.*, 95(450):449–485, 2000.

[4] Venzon, D. & Moolgavkar, S., A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, 37(1):87–94, 1988.

[5] Walter, E., *Identifiability of Parametric Models.* Pergamon Press, 1987.

[6] Meeker, W. & Escobar, L., Teaching about approximate confidence regions based on maximum likelihood estimation. *The American Statistician*, 49(1):48–53, 1995.

[7] Press, W., Teukolsky, S., Flannery, B. & Vetterling, W., *Numerical Recipes: FORTRAN*. Cambridge University Press, 1990.