Validation of systems biology models
Hasdemir Durmus, D.

Citation for published version (APA):
6

Conclusions, Reflections and Perspective

6.1. Analysis of nonlinear kinetic models

6.1.1. Model validation

ODE based kinetic models usually include a large amount of parameters whose values are unknown. A consensus strategy is estimating the unknown parameters from data. The data used for this purpose is typically time series concentration values of biochemical species involved in the model. When a validation step also has to be considered, a pre-determined part of the available data is excluded from the parameter estimation step. The predictive power of the model on the left-out part of the data is used for model validation or selection purposes in the case of multiple competing model structures. In the light of our findings from Chapter 3, we conclude that the partitioning of the data is not a trivial task. Decisions regarding the validity or superiority of a model structure compared to others are affected by the choice of the partitioning scheme. This is because some parts of the data are essential for parameter inference due to reasons originating from the underlying biological properties of the system in question. However, determination of these parts beforehand is not possible since it depends on the parameters of the model which might be unknown. This fundamental drawback associated with the hold-out approach has been underestimated in the field. Stratified random cross validation overcomes the limitations of the hold-out approach by making use of
each part of the data in an iterative and structured manner.

6.1.2. Standards in ODE modeling

Model parameters that have been identified \textit{in vitro} can be accessed via established enzyme databases such as Brenda [131]. However, most kinetic models deposited in model databases such as Biomodels Database [94] contain a substantial number of parameters estimated using \textit{in vivo} concentration data. MIRIAM (Minimum information required in the annotation of models [118]) protocol is an initiative to standardize the curation of quantitative biological models. It sets certain rules primarily regarding the correspondence information of the model and the format in which the model is deposited. One of the requirements entails the usage of a machine readable format such as SBML (Systems Biology Markup Language [67]) which provides a standardized representation of the model in terms of the model constituents and the governing biochemical rules. The optimal parameter values estimated in the associated studies are embedded in the deposited model files and are available to be used by other modelers. However, information about the characteristics of the data such as the experimental conditions under which the data were collected is not provided in a standardized format and is available only in the reference article describing the model. Keeping in mind that independent data are needed for model validation, it is crucial to have detailed information on the data from which the parameters of the model were inferred. It is not difficult to realize that standardizing the presentation of such information is a fairly high target to achieve. However, it is ultimately needed for paving the way to more systematic construction and validation of quantitative models.

6.1.3. Optimal experimental design

There is substantial literature presenting guidelines for optimal experimental design for the analysis of ODE based models [27, 46, 89, 137, 140]. The optimal design is dependent on the ultimate goal of the modeler. The optimal experimental conditions, measurement points and the needed levels of input parameters (such as disturbances to the system) change according to whether the primary goal is better estimation of the parameters or better discrimination of competing model structures [89]. In other words, the experimental conditions that have to be used for parameter inference with less uncertainty are not necessarily the same with those that have to be used for improved selection between alternative model structures.

Our findings in Chapter 3 can be considered as complementary to this fact. In Chapter 3, we have shown that the experimental conditions that are appropriate to
validate a model structure are not necessarily appropriate for model selection. This is because, other competing models might also perform well under these experimental conditions. The concept of optimal experiment design and our cross validation concept agree on the fact that the details regarding the experimental conditions are of vital importance. The application of the optimal targeted experimental design concept is however, almost not existent in the field. Speculating on the reasons; this is partly due to the fact that the optimal conditions for experiments are assumed to be known to biologists through prior biological information on the system. However, efficient construction of reliable quantitative models definitely rely on well designed experiments.

6.1.4. Assessment of predictive power
Models are abstractions of the reality and should be as simple as possible but yet capable of explaining the observed data and of predicting the unobserved data. In Chapter 2, we have shown that cross validation provides a means of calculating the predictive power of kinetic models when applied in a stratified way, leaving out as test set in each turn, sets of data points which are homogeneous in time points and biochemical species. We have also shown that the final assessment of a model’s predictive power’s sufficiency can be made by comparing it to the predictive power of an unsupervised approach which does not rely on any biochemical knowledge. Smooth principal components analysis (SPCA) works well as such an empirical threshold, since it relies not only on the correlation between the species but also on their time dependent smooth behavior.

It is important to re-emphasize that we assess the predictive power of a model by this approach. The task accomplished here is, hence, not serving the same purposes as residual tests for assessing the quality of model fit [29]. In our approach, the data on which the prediction residuals are obtained have to be excluded from the data that are used for parameter inference. For this purpose, we either leave out interior time points or a set of consecutive time points at the end of the complete time profiles of biochemical species. Upon reflecting on our work, we believe that leaving the interior time points out as test set points can be slightly controversial. This is because, data at a certain time point are not fully independent of the neighboring time points. Therefore, one can argue that when a time point is left out of the training set, the neighboring time points are enough to represent it in the training set. However, we see clear distinctions between the predicted time profiles when different time points are left out of the training set. This proves that removing interior time points leads to the loss of information that they carry even though the neighboring points are still in the training set.
Excluding consecutive time points at the end of the profiles from the training set can be considered as a better approach regarding the same dependency problem. Since forecast analysis is about making predictions regarding unobserved outcomes in the future, it does not suffer from the problem of dependency between the time points. It is also more appealing from a biological point of view. However, this approach has to be applied with care since test sets should not be composed of only the time points that are in the steady-state region. Otherwise, the technique would lead to trivial test sets.

In this respect, leaving out consecutive interior time points (windows of data) as test sets can be more satisfactory [102, 134] for proper application of cross validation on time series data. In this way, when windows of data are left out, the remaining time points would be less representative of the left out time points.

An exceptional case where our approach might not be directly applicable would be a quantitative model consisting of biochemical species with a significantly lagged correlation structure. If the time needed for them to arrive at steady state differs significantly, problems can occur regarding the application of SPCA which depends on the correlation. Therefore, simulations on toy models of such a structure would help to test the applicability of our approach in such cases.

Another possible point of improvement is due to the typical unidentifiability problem of ODE based models. Due to scarcity of the data and high amount of unknown parameters, many of the parameters in such models can only be estimated with very large confidence intervals [125, 129]. The average predictive power of the model across the whole set of parameters in the confidence interval can be, therefore, claimed as a better indicator of the specific predictive power of a model structure instead of the power calculated only at the optimal parameter setting. Such an approach requires a good definition of the confidence intervals of parameter estimates obtained by e.g. bootstrapping approaches [76, 82] or posterior distributions of parameters obtained by a Bayesian perspective [155].

The stability of our approach can be improved by using bootstrapping. Our results from Chapter 2 indicate that model invalidation decisions are more affected by the idiosyncratic noise realization in the data when the experimental noise is high. Complementing our approach with bootstrapping would help to decrease the effect of noise on the decisions when high levels of experimental noise in the data are suspected. In this approach, prediction errors have to be calculated as an average of all bootstrap samples.

As a final remark, we should stress once more that our approach assesses the predictive power of a biochemical model structure based on a given dataset. Therefore, the decisions can change depending upon the dataset. For example, a previously
validated model structure might be invalidated when a dataset with more measurement points in time becomes available. However, this is valid for any validation approach that exploits the consistency between the model predictions and the data.

6.2. Transcriptional regulatory networks

In Chapter 4, we have exploited network component analysis to detect inconsistencies between gene expression data and fixed topology of a transcriptional regulatory network. Discrimination of competing static network models is possible by looking at the measures of consistency. Local improvements of existing topology can also be made through the application of our approach by identifying missing connections. The last feature is valid even when the experimental noise and the number of misconnections and missing connections in the data are fairly high as would be in reality. Furthermore, the local improvement feature of our approach can be extended to sets of genes which were not initially in the proposed network structure. These make it possible to apply our approach iteratively in a modeling-experiment cycle. Potential points of improvement in the networks can be detected by computational means and can be confirmed by targeted experiments.

Our approach, however, might suffer from the limitations of the network component analysis [99] approach upon which it is built. A primary criterion for the applicability of network component analysis is that the transcription factor activity matrix has to be full rank. We conclude that this can be disturbed when the activities of transcription factors are heavily dependent on each other. This problem can be solved by exploiting data from a larger number of experimental conditions. However, this can also result in instability of the transcription factor - gene connection strengths. The experimental conditions used should not be very distinct from each other to keep the connection strength values constant across these conditions. Due to this reason, the selection of the experimental conditions used in such an analysis is a nontrivial task.

6.3. Clustering of large scale biological data

6.3.1. Assessment of validity

Every clustering algorithm gives its own view of the data. Therefore, validating the results from a specific clustering algorithm is necessary. Stability of the resulting clusters is an important feature of the cluster analysis. High stability against small perturbations in the data which are due to the specific realization of the experimental noise shows the reproducibility of the clusters. Therefore, stability provides a means
of validating the analysis. However, it is important to investigate the behavior of the stability measures on the original data relative to a reference value. Analytical calculation of such a reference value would be highly demanding in terms of the assumptions needed for describing the distribution of the data. For this purpose, we have employed a rather different approach in Chapter 5. Observing the stability level obtained from the clustering of a densely distributed, single cluster synthetic dataset helps us to have an understanding of how stable clustering on data without a nontrivial cluster structure would be. We have applied this comparative technique in its very crude form, though.

The stability of the single cluster dataset might be affected by factors such as the size of the dataset, amount and structure of the experimental noise, and the level of covariance between the variables. At relatively high noise levels, the specific realization of the noise can be influential by itself. Therefore, the issue of setting a well defined reference point deserves a detailed study. Simulations by using datasets differing at the levels of the aforementioned factors and considering their average stability level would be beneficial for this purpose.

6.3.2. Parameter optimization in cluster analysis

A major task in a cluster analysis is the optimization of the parameters of a clustering algorithm. In many of the algorithms, the number of clusters imposed is a parameter that can be optimized by using the data. Stability measures employing cross validation [147] and quality measures such as the distance based metric Silhouette Width [126] have been proposed for this task. However, when the clusters in the data are not well separable, the optimal number of clusters obtained by these measures are very low. This leads to very coarse clustering of the data which is too rough for biological interpretation. Optimality might have to be traded off to gain more biological interpretability. Therefore, we need measures that regulate this trade-off. The metric introduced in [52] and the biological homogeneity index of [34] both score the biological relevance of the cluster analysis based on the functional categories to which the genes belong. The biological stability index of [24, 34] takes it one step further and helps in the stability assessment of the co-clustering of functionally related genes. However, it does not specifically regulate the trade-off we have mentioned.

6.3.3. Dealing with vague structures

The clusters in biological data are often not well separable. It is not hard to imagine this phenomenon since we are dealing with overlapping groups of genes that belong
to multiple functional categories. In this manner, fuzzy clustering algorithms provide a suitable framework. However, they require more complex algorithms compared to their traditional counterparts. Consequently, their validation and the optimization of their parameters still deserve active research [164].

Application of different clustering algorithms in parallel is also promising to deal with vaguely structured datasets. Detecting the commonly emerging patterns in the data by multiple approaches is important in the sense that it paves the way to consensus clusters that are free of the individual bias of the algorithms. However, patterns that can only be detected by particular algorithms might also be very important since not every algorithm can reveal different subtle structures in the data. Therefore, not only the comparison but also the reliable integration of different clustering algorithms should attract attention in the field of clustering analysis.

6.3.4. Incorporation of validation and quality measures in cluster analysis software
Software packages are available to perform external validation and to calculate internal quality measures, in most common languages for statistical analyses such as Matlab [162] and R [24]. However, rather dedicated, standalone software for cluster analysis are extensively used by biologists. Validation of clustering has to be conceptually more integrated in the cluster analysis and thus validity measures have to be incorporated also in standalone cluster analysis software.

6.4. Incorporating resampling approaches
6.4.1. Error models for bootstrapping
Applications of bootstrapping proves to be beneficial in the analysis of uncertainty of models and detecting model stability in a variety of systems biology applications including clustering and nonlinear kinetic models of biochemical systems. However, reliable conclusions can be achieved only by employing reliable bootstrap datasets. Reliability of a bootstrap sample depends on how well repeated sampling of the data is mimicked by the bootstrap samples. There are certain assumptions in constructing bootstrap samples and hence, the reliability depends on how realistic these assumptions are. These assumptions are related to the error structure in the original data such as the exchangeability of the experimental error term between the time points or between the biochemical species in the model. Their fulfillment depends on the experimental methods used to collect the original data. Therefore, resampling approaches can prevail more in systems biology if studies focus on modeling
the error structures associated with individual types of experimental methods.