Learning Process Models with Missing Data

Will Bridewell, Pat Langley, Steve Racunas, and Stuart Borrett

Computational Learning Laboratory, CSLI,
Stanford University, Stanford, CA 94305 USA
{willb,langley,sracunas,sborrett}@csli.stanford.edu

Abstract. In this paper, we review the task of inductive process modeling, which uses domain knowledge to compose explanatory models of continuous dynamic systems. Next we discuss approaches to learning with missing values in time series, noting that these efforts are typically applied for descriptive modeling tasks that use little background knowledge. We also point out that these methods assume that data are missing at random—a condition that may not hold in scientific domains. Using experiments with synthetic and natural data, we compare an expectation maximization approach with one that simply ignores the missing data. Results indicate that expectation maximization leads to more accurate models in most cases, even though its basic assumptions are unmet. We conclude by discussing the implications of our findings along with directions for future work.

1 Introduction

Consider the challenge of collecting ecological data from the Southern Ocean. The location is remote, the climate can be brutal, and scientists have limited resources, which forces them to carefully plan and prioritize their collection efforts. To accomplish this task, scientists schedule observation cruises when and where they anticipate that phenomena of primary interest will occur. However, the spatial and temporal variability of ecological phenomena further hampers data collection. The phenomena may not occur where anticipated, may happen before or after a cruise, or may last longer than a single research cruise can remain at sea. One strategy to address this issue is to make multiple cruises, but even if this approach is successful, there will be omissions in the data. Yet scientists still want to build models and determine parameter values to explain the data and understand the system.

This scenario highlights a number of issues. First, the gathered data will likely contain large gaps that were engineered from the start. Second, these gaps depend partly on the expected values of the missing data—they are not missing at random. Third, even though the gathering efforts are engineered to contain the most important information, the timing may be off. In addition, instruments may malfunction and some environmental values may fall outside measurable ranges. Despite the scientists' best efforts, important information about system dynamics may be missing.

All these situations, which are not unique to large ecological expeditions, leave the scientist with an interesting and complex problem: how can one build

a model of a nonlinear, dynamic system when key measurements are missing? Inductive process modeling (Langley et al. 2002) provides a method for building quantitative explanations from time series, but it assumes that the relevant data are available. In comparison, ARIMA methods lead to purely descriptive models, but researchers often augment them with a variant of expectation maximization (EM; Dempster et al. 1977) to handle missing values (Isaksson 1991; Stoica et al. 2005). In this paper, we determine whether EM can be adapted to assist inductive process modeling and to function in realistic scientific settings.

In the pages that follow, we apply an EM variant called EMP to produce an explanatory model of scientific data and compare its behavior to a baseline method that ignores the missing values. The next section describes the inductive process modeling paradigm and introduces the baseline and EM approaches to handling missing observations. After this, we report experiments with synthetic and natural data and present an analysis of the results. In closing, we discuss related work and suggest directions for future research.

2 Handling Missing Data in Inductive Process Modeling

The approach we report here extends earlier work on inductive process modeling (Langley et al. 2002; Todorovski et al. 2005). The discovery task can be stated:

- *Given*: trajectories for a set of continuous variables over time;
- *Given*: background knowledge cast in terms of generic entities and processes;
- *Given*: observable and theoretical entities and variables to be modeled;
- *Find*: a process model that explains the observed trajectories and generalizes accurately to new data.

The task revolves around the notion of a quantitative process model, which provides a causal account of how variables change over time. Todorovski et al. (2005) describe the process model representation, which consists of distinct processes that organize numeric relations among variables that are associated with known entities, and introduce HIPM, a program that induces process models.

Inductive process modeling should lead to a mathematical model of a system that both improves our understanding of that system and enables us to predict its behavior under altered conditions. Additionally, one could use the model to reconstruct unobserved points in a set of trajectories—a use that suggests a solution for handling missing data. For instance, given a model, we could substitute its output for the missing values back into the original data set and learn a new model from the result. This approach falls into the expectation maximization (EM) class of techniques (Dempster et al. 1977).

The EM algorithm is an iterative approach to learning a model from data with missing values that has four main steps:

1. select an initial set of parameters for a model
2. determine the expected values for the missing data
3. induce new model parameters from the union of the expected values and the original data
4. unless the parameters have converged, return to Step 2 using the new model

To be applicable without further complication, EM assumes that the mechanism responsible for the missing data is ignorable. Specifically, the data must be either *missing completely at random*, which means that the mechanism is independent of the observations, or *missing at random*, where the data may influence it (Little and Rubin 2002). Unfortunately, scientific data sets rarely meet these criteria.

Missing values in scientific domains arise from a combination of resource constraints, working hypotheses, and other reasons both practical and accidental. Although variants of EM exist for such "non-ignorable" mechanisms, they require a collection of data sets produced with the same missing-data mechanism—a luxury not typical to scientific research. In response, we chose to violate the assumption of an ignorable mechanism and experimentally evaluate the utility of an EM variant under these conditions. Our specific algorithm, which we call EMP, follows the general EM outline given earlier:

1. substitute linearly-interpolated values for the missing data
2. use HIPM to find the model that minimizes the sum of squared errors
3. simulate the new model
4. substitute the results of the simulation for the missing data
5. if the model has changed and the maximum number of iterations has not been exceeded, go to Step 2

This algorithm differs from the previous outline in that the missing data are initially replaced with rough estimates and that HIPM selects the first model as well as the subsequent ones. In addition, estimating parameters for a nonlinear system remains an unsolved problem, and we can guarantee neither a maximization nor an improved estimate in Step 2. Thus we introduce a maximum number of iterations to force the program to halt. In the next section, we compare the results from EMP to those from a baseline that ignores the missing data.

3 Experimental Evaluation

In the last section, we established that EM is not an ideal fit for the missing data problem that we encounter, but we also saw that a variant of EM may be a reasonable solution in practice. To test this conjecture, we performed experiments with synthetic and natural data for a two population predator–prey system and with natural data from a more complex ecosystem. The synthetic case allows us to control the nature of the noise in the trajectory and to determine how accurately HIPM can recreate the structural and parametric form of the generating model. The natural data gives a more realistic view of EM's capabilities in the presence of complicated and unknown noise models.

To generate synthetic data, we built a predator–prey model that produces a stable oscillation as would be expected in an ideal system. This model includes a process for logistic growth of the prey, exponential death of the predator, and a Holling type 2 function (Holling 1959) for predation. We simulated the model to mimic experimental conditions where four measurements are taken each day for 35 days, which gave a total of 141 observations including the initial conditions.

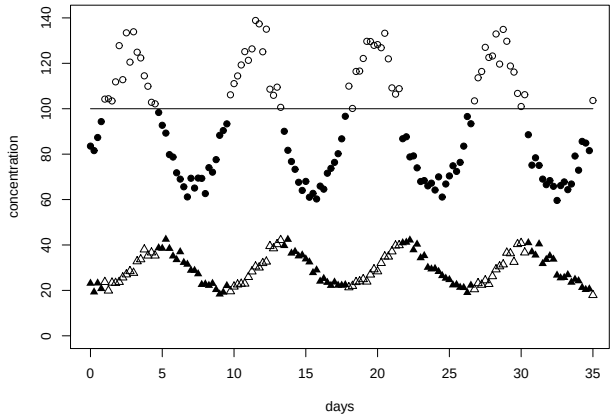

Fig. 1. This figure shows the synthetic predator–prey data used in the experiments. The horizontal line denotes the cut point for the peaks in the prey concentration. Unfilled shapes indicate missing values.

To generate the final trajectories, we added five percent multiplicative, Gaussian noise to each observation.

After generating the data, we altered the trajectories to produce a plausible worst-case scenario for model induction. For this experiment, we assumed that peaks in the prey population were of primary importance, so we removed them by deleting all observations where the concentration of prey exceeded 100 parts/volume. This operation left roughly half of the data for training purposes, as shown in Figure 1. For the baseline condition, HIPM searched exhaustively through a space of 22 model structures to fit the corrupted training data and tested the resulting model on the original, noise-free trajectories. We used the same search conditions to test EMP, which performed 20 iterations and reported the model with the lowest sum of squared error over all the iterations.

To evaluate EMP on observations from a real system, we used two data sets initially collected by Veilleux (1976) and made available by Jost and Ellner (2000). In his experiments, Veilleux observed the interaction of two protist species in an artificial environment over several days. Since it typically took a few days for these ecosystems to establish a stable frequency, we use a subset of the provided values. Specifically, we use the observations shown in Jost and Ellner's Figure 1a starting at day 8.5 and those in their figure 1c starting at day 11. The resulting data sets contain 54 samples with five full peaks and 30 samples with three full peaks, respectively.

As with the synthetic data, we removed the portions of the Veilleux trajectories that contain the prey's peak values. Here we tested under two conditions. In the first case, we cut out the peak value and one or two neighboring points, slightly shaving off the peak. For the second case, we removed the peak value and three to four surrounding points, imposing a deeper cut. Note that a full cycle, from trough to trough, contains ten samples on average, so the second scenario uses roughly half of the total data. HIPM fit each data set independently by searching the same 22 structures used with the synthetic data. We carried out

Table 1. Results on synthetic and natural predator–prey data. The mean squared error (best scores in bold) and coefficient of determination (r^2) are reported for the best models produced by the baseline approach and EMP.

Data	Mean Squared Error		Predator r^2		Prey r^2	
	Base	EMP	Base	EMP	Base	EMP
synthetic	44.04	**13.34**	0.88	0.97	0.89	0.95
1a minor	2073.93	**1925.28**	0.65	0.64	0.63	0.62
1a major	**2580.81**	2636.63	0.55	0.58	0.54	0.55
1c minor	245.42	**231.98**	0.87	0.87	0.91	0.90
1c major	408.67	**249.11**	0.88	0.87	0.90	0.90

the experiments with EMP in the manner previously described and measured each model's accuracy by testing it against the original, uncut data.

Table 1 shows the results for the predator–prey experiments. In all but one case, EMP produced models with substantially better fits to the original trajectories than did the baseline approach. Interestingly, neither method reproduced the correct model structure for the synthetic data, although HIPM can recover it from perfect data. Notice also that the coefficients of determination for both variables (r^2) are roughly the same across methods. This result suggests that EMP helps HIPM fit the amplitude of the trajectories, but does not affect its ability to fit their shapes. Plots of the trajectories, such as the one in Figure 2, show that, in all cases, the corresponding models provided close visual fits to the frequency in both the 1a and 1c data sets. In addition, the models fit the amplitude of the 1c data quite well, but produced peaks roughly half the height of those observed in the 1a data.

The final experiments evaluate our approach with data from the Ross Sea in the Southern Ocean. This domain differs from the previous two in that the space of model structures is much larger and the data contain a single peak in the primary variable. For the experiments we used two data sets (RS1 and RS2) provided by Kevin Arrigo, the oceanographer in our group. Each set contains 188 preprocessed, daily observations of phytoplankton and nitrate concentrations, along with values for the amount of available light. In both cases the recordings were made over the summer when a single phytoplankton bloom occurred.

We removed 32 samples from the first year of data and 25 samples from the second, based on the location of the phytoplankton peaks. Since light serves as a driving variable, we provided its value in all cases. After preparing the data, we had the program fit each set independently and compare the results against the original measurements. Due to the size of the search space defined by the associated generic process library, we ran HIPM in beam-search mode with a beam width of eight. For each training cycle, the program considered an average of 126.7 model structures. As before, EMP ran a total of 20 cycles using the same settings for HIPM in all cases and returned the model with the lowest sum of squared errors.

Table 2 shows the results on the Ross Sea data. In both cases, EMP substantially outperforms the baseline in terms of both mean squared error and r^2. We

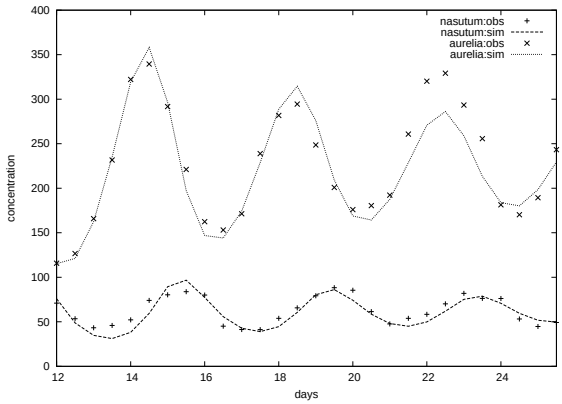

Fig. 2. This figure presents the trajectories produced by EMP's best model for the Veillieux 1c data along with the observed values.

see more than a 50% reduction in error and, although the r^2 for phytoplankton decreases a bit, the increase for nitrate is phenomenal. Without the use of EMP, HIPM predicted a flat line for the nitrate concentration. In summary, EMP not only reduced error but also improved the conceptual model by accounting for all the observed variables.

The results presented above paint a highly positive picture of the EM approach to handling missing data in inductive process modeling. On both the synthetic and Ross Sea data, the extra computational time led to much better fits, whereas the fits on the Veilleux data were mostly improvements. In the next section, we review the experimental results, suggest further work in this area, and discuss related research.

4 Discussion

Even though EMP is hampered by an unspecified missing data mechanism and asked to operate in a worst-case scenario, it behaved quite well. Looking more closely at the results on the Veilleux data set, we can conjecture why EM was less helpful in some cases and plan studies that could clarify the reasons. First, we note that the behavior of both EMP and the baseline on the 1c data matches what we see when HIPM induces a model from the complete data. Thus, EMP was likely hitting a performance ceiling, and enough information remained in the data for HIPM to build an accurate model even in the baseline condition. This result is somewhat surprising, since over one-third of the data were removed. Second, we could make a similar argument for the 1a data, but performance of both approaches degrades when we remove more of the data. This finding matches intuition and indicates that we corrupted the data enough to affect HIPM's performance.

We should also explain why experiments with the synthetic predator–prey data highlighted the difference between EMP and the baseline more clearly than those using the Veilleux sets. The most obvious difference in these two cases is the

Table 2. Results on data from the Ross Sea. The mean squared error (best scores in bold) and coefficient of determination (r^2) are reported for the best models produced by the baseline approach and EMP.

Data	Mean Squared Error		Phytoplankton r^2		Nitrate r^2	
	Base	EMP	Base	EMP	Base	EMP
RS1	30.13	**13.56**	0.98	0.96	0.00	0.87
RS2	25.27	**9.93**	0.93	0.80	0.00	0.93

nature of the noise in the observations. The synthetic data used a multiplicative, Gaussian model whereas the noise mechanism of the natural data is unknown. Further experiments with alternative noise models may help clarify the effect of noise on both methods and determine whether it accounts for the discrepancies seen in the results.

Although our approach to the missing data problem is related to previous techniques, we have adapted it to the inductive process modeling task and examined its ability to work on scientific data. Process modeling, which we described earlier in the paper, descends from research on equation discovery (e.g., Langley 1981; Żytkow et al. 1990; Džeroski and Todorovski 1995; Washio et al. 2000), but it differs in that it takes background knowledge as input and outputs explanatory models, as opposed to descriptive ones. Our emphasis on differential equations bears some resemblance to work by Bradley and colleagues (2001) and Todorovski (2003), but these approaches lack a a strong attachment to scientifically meaningful processes.

We could also characterize inductive process modeling as a combination of qualitative physics and system identification. In particular, our approach groups equations into a more qualitative, process-based structure like that developed by Forbus (1984), and our use of generic processes to encode background knowledge resembles work in compositional modeling (e.g., Falkenhainer and Forbus 1991), where abstract components are instantiated and assembled to form models. The relationship to system identification (Åström and Eykhoff 1971) lies in our concern with learning parametric models from time series. Inductive process modeling differs from this paradigm in its incorporation of search through a space of model structures.

Although this paper indicates that EM is an appropriate technique for handling missing data when learning process models, more work in this area remains. Here, we concentrated on the case where large portions of data are unavailable, but other situations often arise. In some cases, the variables may be measured at different intervals, which in extreme cases results in a collection of examples that are missing all but one value. In other cases, certain variables may be recorded only at specific times, as occurs when data sets are merged from multiple sources, each reflecting different interests and resource constraints. This situation can cause large gaps in individual variables without affecting the rest of the data. We conjecture that EM-style techniques will be useful in these situations, but we need experiments to test this prediction.

Acknowledgements

This research was supported by Grant No. IIS-0326059 from the National Science Foundation. We thank Ljupčo Todorovski for design and implementation of the HIPM system and Oren Shiran for extensions to the software.

References

Åström, K. J., Eykhoff, P.: System identification—a survey. Automatica **7** (1971) 123–167

Bradley, E., Easley, M., Stolle, R.: Reasoning about nonlinear system identification. Artificial Intelligence **133** (2001) 139–188

Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) **39** (1977) 1–38

Džeroski, S., Todorovski, L.: Discovering dynamics: From inductive logic programming to machine discovery. Journal of Intelligent Information Systems **4** (1995) 89–108

Falkenhainer, B., Forbus, K. D.: Compositional modeling: Finding the right model for the job. Artificial Intelligence **51** (1991) 95–143

Forbus, K.: Qualitative process theory. Artificial Intelligence **24** (1984) 85–168

Holling, C. S.: Some characteristics of simple types of predation and parasitism. Canadian Entomologist **91** (1959) 385–398

Isaksson, A.: System identification subject to missing data. American Control Conference (1991) 693–698

Jost, C., Ellner, S.: Testing for predator dependence in predator-prey dynamics: A non-parametric approach. Proceedings of the Royal Society of London B **267** (2000) 1611–1620

Langley, P.: Data-driven discovery of physical laws. Cognitive Science **5** (1981) 31–54

Langley, P., Sánchez, J., Todorovski, L., Džeroski, S.: Inducing process models from continuous data. The Nineteenth International Conference on Machine Learning (2002) 347–354

Little, R. J., Rubin, D. B.: Statistical analysis with missing data (2nd ed.). Hoboken, NJ: John Wiley

Stoica, P., Xu, L., Li, J.: Parameter estimation with missing data via equalization-maximization. IEEE International Conference on Acoustics, Speech, and Signal Processing (2005) IV–57–IV–60

Todorovski, L.: Using domain knowledge for automated modeling of dynamic systems with equation discovery Doctoral dissertation, Faculty of computer and information science, University of Ljubljana (2003)

Todorovski, L., Bridewell, W., Shiran, O., Langley, P.: Inducing hierarchical process models in dynamic domains. The Twentieth National Conference on Artificial Intelligence (2005) 892–897

Veilleux, B. G.: The analysis of a predatory interaction between didinium and paramecium. Master's thesis, University of Alberta (1976)

Washio, T., Motoda, H., Niwa, Y.: Enhancing the plausibility of law equation discovery. The Seventeenth International Conference on Machine Learning (2000) 1127–1134

Żytkow, J. M., Zhu, J., Hussam, A.: Automated discovery in a chemistry laboratory. The Eighth National Conference on Artificial Intelligence (1990) 889–894