Analysis of Sibling Time Series Data: Alignment and Difference Detection

by

Jennifer Listgarten

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

Abstract

Analysis of Sibling Time Series Data: Alignment and Difference Detection

Jennifer Listgarten
Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto
2007

Many practical problems over a wide range of domains require synthesizing information from time series data. Two distinct, yet related, problems in time series data are those of alignment and difference detection. These tasks may be coupled together so that a solution to one is difficult without a solution to the other.

We introduce a unified, probabilistic approach to the problems of alignment and of alignment with difference detection. This approach takes the form of a class of models called *Continuous Profile Models* for simultaneously analyzing sets of sibling time series – those which contain shared sub-structure, but which may also differ. In this type of generative model, each time series belonging to one class is generated as a noisy transformation of a single *latent trace* in the model. A latent trace can be viewed as an underlying, noiseless representation of the set of observable time series belonging to one class, and is learned from the data. If multiple classes of data exist, then one latent trace per class is learned, and these are aligned to each other during inference. The latent traces lie at the core of this class of models, and provide the basis for alignment and difference detection.

Our approach to alignment has several benefits over traditional approaches. It provides a principled method for finding parameters in the model, such as the reference template and error/distance function, rather than specifying these in an *a priori* and/or *ad hoc* manner. It simultaneously aligns all data in one go, rather than aligning them in a greedy, incremental fashion. It corrects scaling of signal intensity while performing alignment. Additionally, the probabilistic framework allows us the option of using fully Bayesian inference, if desired, so

that we may gauge uncertainty in our model parameters, integrate out model parameters, and avoid cross-validation, which can be problematic with limited data. Lastly, the CPM is the first model, to our knowledge, to tackle the simultaneous problems of alignment and difference detection.

We focus on Liquid-Chromatography-Mass Spectrometry proteomics data, for examination and demonstration of our methods, although our methods are not confined to this domain.

Acknowledgements

By far the most important acknowledgment and thanks goes to my two supervisors, Sam Roweis and Radford Neal. Each one is unsurpassed as a supervisor, except by their combination. Additionally, the machine learning group in Toronto – faculty, postdocs and students alike provided a wonderful environment in every respect. In particular I would like to thank Roland Memisevic, Rama Natarajan, David Ross, Ted Meeds, Nati Srebro, Ben Marlin and (honorary ML member due to his use of our lounge) Horst Samulowitz. Also, I would like to thank Anna Bretscher, Josh Buresh-Oppenheim and Diego Macrini for their long and lasting friendship throughout and past our studies together in Toronto, Kathryn Graham for her biological knowledge and Alberta beer, and Jack Newton for his endless supply of technical knowledge, gourmet cooking and parties. I thank my parents for their constant support. And JR for all.

I would also like to thank my other committee members, Brendan Frey and Andrew Emili, and my external examiner, Kevin Murphy for their valuable feedback, and also Ben Marlin for his slice sampling function, Tom Minka for his Matlab Lightspeed and Fastfit toolboxes, Carl Rasmussen for his Matlab minimize routine, and Giorgio Tomasi for his Matlab implementations of DTW and COW.

Part of my funding came from the Natural Sciences and Engineering Research Council of Canada and the Ontario Graduate Scholarship program.

Contents

Chapter 1

Continuous Profile Models

1.1 The Alignment Problem

Many practical problems in many domains require synthesizing information from data sampled over time or space. An example of such *time series* are speech waveforms, in which the digital representation of a vocal utterance is represented by real-valued numbers at a sequence of discrete time points. Time series data are often noisy, and in particular, the timing of salient events can be extremely variable. This variability arises because timing during collection of the data cannot be accurately controlled, or, if it can be controlled, the measured time does not correspond to the timing of the underlying processes we wish to model and understand. For example, speech recognition needs to account for the fact that different people speak at different rates. The underlying processes producing speech are the utterances of particular syllables (or phonemes), which can span more or less time, depending on the particular speaker. It is useful to find some canonical/reference time frame to which all the time series can be mapped, so as to make them directly comparable to one another. We refer to the problem of making two or more time series comparable to one another by changing their relative timing as the *alignment* problem.

It is valuable to note that the alignment problem is ill-posed. That is, there does not exist a single, best solution which could be agreed upon. If one has speech data from several speakers, say with each speaking at his/her own uniform rate, then what would be the best reference time frame to which to map all of these speech time series? One could just as well use the slowest person's time frame, or the fastest person's, or any of the ones in between. Or one could use a time frame that was never observed. One could even devise a pathological time frame in which the original data was mapped non-monotonically (*i.e.,* the relative timing of events in a given

time series would change so that the future with respect to one event becomes the past in the pathological time frame).

Are any of these reference frames better than the others? One could argue quite convincingly that the pathological time frame is not desirable because we would like a reference time frame which maps events monotonically. But beyond this criterion, can we reasonably argue for one time frame over another? One could appeal to the Occam's razor principle which states that simplicity is best – all else being equal. In the alignment setting, this principle could be translated into the desideratum that a reference time frame should be one in which the least 'complex' mappings from observed time series to the reference time is achieved, provided the quality of the alignment is not sacrificed (the best quality alignment would be one in which the underlying processes are in perfect correspondence with each other). Indeed, adherence to Occam's principle would rule out our pathological reference frame since swapping past and future is surely a complex mapping by any definition. This idea of keeping it simple enters into traditional alignment algorithms, under the guise of 'regularization', in which model complexity is held at bay.

The alignment problem is a pervasive one, spanning not only speech and music processing, but also, for example, equipment and industrial plant diagnosis/monitoring, and analysis of biological time series from microarray and liquid/gas chromatography-based laboratory (LC/GC) data (such as mass spectrometry (LC-MS) and ultraviolet diode arrays). A main contribution of this thesis is to provide a robust model and algorithm for alignment of time series data.

1.2 Amplitude Normalization

In addition to the need for correction of time, one may also need to correct for systematic differences in the amplitude of the signal at each time point. This is the problem of *normalization*. For example, in a laboratory experiment, it may be the case that an instrument is slightly miscalibrated one day, and that all measurements are thus systematically larger than on a previous day. Or it may be the case that some speaker's volume trails off at the end of a sentence, while others' remains constant, or takes on some other pattern. These are not variations in the timing of events, but in the amplitude of events at each time point. Normalization might be viewed as an easier, or less critical problem than alignment, although the two can be intertwined – that is, one may only be able to do 'optimal' alignment after normalization, and one may only be able to do 'optimal' normalization after alignment. Indeed, there is a degeneracy arising from performing both of these corrections, since one could imagine putting an observation in one

time series into corespondance with that in another either by moving it in time, or by changing its scale.

1.3 Alignment of Liquid-Chromatography Experiments

As mentioned, the problem of alignment arises in many biological or chemical experimental settings, such as experiments which measure the expression of genes at various points in the cell cycle (using microarrays). In this thesis, liquid-chromatography based experiments are the primary application. In liquid chromatography, some solution of interest (*e.g.*, serum) is passed through a chromatography column which separates parts of the solution on the basis of some chemical property (for example, hydrophobicity). At discrete time intervals, solution is collected as it exits the column, and then is analyzed by an instrument (*e.g.*, a mass spectrometer). Analysis by the instrument at each discrete time point can provide a feature vector (*e.g.*, mass/charge ratios in mass spectrometry), or scalar values (*e.g.*, UV absorbance at a single wavelength).

If a single specimen of solution is split into two parts, and each of these are then run through the same liquid chromatography column, they will not travel through the column in identical ways. Their paths are affected by chemical and physical properties of the column which may not remain identical from run to run or even within a run, as well as by ambient factors such as temperature and pressure in the laboratory. Thus data collected from LC experiments can be extremely variable in time.

Liquid-chromatography (LC) techniques are currently being developed and refined with the aim of providing a robust platform with which to detect differences in biological organisms – be they plants, animals or humans. Detected differences can reveal new fundamental biological insights, or can be applied in more clinical settings. LC-mass spectrometry technology has recently undergone explosive growth in tackling the problem of biomarker discovery – for example, detecting biological markers that can predict treatment outcome or severity of disease, thereby providing the potential for improved health care and better understanding of the mechanisms of drug and disease. In botany, LC-UV data is used to help understand the uptake and metabolism of compounds in plants by looking for differences across experimental conditions.

Many of the models and algorithms introduced in this thesis are examined in the context of LC data, although our models and algorithms are very general, and could of course be applied to any number of domains.

1.4 The Difference Detection Problem

In many time series settings, one is interested in aligning data for the purpose of comparison. Comparison could take the form of a simple yes/no response. For example, *does a particular speech utterance match a stored record* – for the purpose of, say, biometric identification. Alternatively, comparison could take on a more detailed form – *what are all of the differences between LC-MS profiles from aggressive prostate cancer serum samples and those arising from a more benign form of the disease.* The first type of comparison, *classification*, is an easier task, since it requires only that we be able to model sufficient information so as to reliably distinguish between two categories. The second type of comparison, *difference detection*, is more challenging since it requires finding all patterns which systematically differ between categories, even if these are rare, noisy or of little value for classification (*e.g.*, are redundant given other information).

Note that there are different ways in which one could define the problem of difference detection. For example, one could define it as looking for all single features which appear in Category A, but not in Category B. Alternatively, one might define it as looking for all feature combinations (*i.e.*, including sets of features) which systematically differ between categories. In other words, there may be individual features which do not differ between categories, but perhaps two such features in combination with one another do provide a systematically different pattern. Technically, we are distinguishing between first order difference detection (using single features), and higher order difference detection (in which *sets* of features are sought).[1] In this thesis, we consider only the first order difference detetecion problem, which is by far the easier problem, since looking for all combinations of features is intractable.

Difference detection manifests itself in many of the same domains as alignment (*e.g.*, biometric speech analysis, industrial plant diagnosis/monitoring, LC-based biomarker discover). When the difference detection problem arises in static data (*i.e.*, not time series data), one can appeal to traditional statistical or machine learning methods to model any systematic differences. However, when one wants to perform difference detection on time series data requiring alignment, the problem can become much more difficult. Difficulty arises if the systematic differences which exist between categories make it difficult to align the data, producing a chicken/egg scenario – to find differences between categories, one may need to first align the data, but to align the data, one may first need to know where the differences are so that they do not disturb the alignment.

[1] Analogously, classification is routinely performed using first order features, or higher order features, or both.

Previous approaches to this problem ignore this chicken-egg scenario, choosing to start with the alignment problem (while ignoring differences between time series), followed by difference detection. That is, people typically apply a *single class* alignment algorithm to the data, assuming that differences need not be accounted for, and then perform difference detection following alignment. A main contribution of this thesis is to present a unified solution to these two, simultaneous problems.

We use the term *sibling time series* to refer to sets of time series which share common substructure, but may also differ from one another. Thus difference detection (and alignment) operate on sibling time series data.

1.5 Contributions of this Thesis

In this thesis, we develop a novel, unified, probabilistic framework for alignment, normalization, and difference detection in sibling time series. In particular, we present a new class of probabilistic, generative models, *Continuous Profile Models* (CPMs), in analogy to Profile HMMs for discrete sequences, which simultaneously *align* and *normalize* sibling time series data. If more than one class of data is present, one can additionally, simultaneously, perform *difference detection*.

The core foundation of Continuous Profile Models is the concept of a *latent trace*, which is a latent variable in these models.[2] A latent trace can be viewed as a canonical, underlying representation of a set of time series data from one class, and is inferred by our algorithms during training/inference. By inferring the latent trace, one can obtain alignments and normalize the data. In the multi-class models, there is one latent trace per class, which provides the foundation for multi-class alignment and difference detection. In this scenario, not only are the latent traces from each class learned, but they are also simultaneously aligned to each other while accounting for differences between them. These class-specific latent traces are then in correspondence with one another and can then be used to reveal differences between classes.

Continuous Profile Models use the formalisms and machinery of probabilistic, generative modeling. This allows for a principled and unified approach to both alignment alone and alignment together with detection. By principled, we mean that we avoid requirements of specifying a reference time frame and a distance/error function for measuring how similar a set of

[2]Latent variables are not directly observed, but explain commonalities among things that are observed. Also, technically, in the first model we introduce, the latent trace might better be considered a parameter, since we are learning only a point-estimate of it.

observed time series are to one another in an *ad hoc* manner. Instead, we use parameters/latent variables which represent these quantities, and learn them from the data in a statistically sound way, thereby providing an algorithm tailored to the data at hand. CPMs are applicable to the alignment problem alone, and also to the problem of simultaneous alignment and difference detection. We emphasize both of these problems equally in this thesis.

Within the CPM framework,

1. We introduce a MAP (maximum *a posteriori*) -based Continuous Profile Model (EM-CPM) for alignment of (single-class) sibling time series data, and investigate properties of this model and algorithm through experimentation on LC-MS data. Training in this model is done using the EM algorithm.

2. We introduce a fully Bayesian, Hierarchical Continuous Profile Model (HB-CPM) which uses Markov Chain Monte Carlo (MCMC) for inference. This model is particularly well suited to performing simultaneous alignment and detection under a single, unified modeling paradigm. We investigate properties of this model and algorithm on some NASA solenoid valve data (providing a nice illustrative example) and on LC-UV data from a botany laboratory.

3. We apply our models to an LC-MS proteomic spike-in experiment in which known proteins are added to a base of serum, providing a realistic, yet verifiable set-up to assess different algorithms. We demonstrate the advantages of using CPMs over the more traditional Dynamic Time Warping class of algorithms for alignment. In this setting, we are also able to demonstrate that the problem of difference detection is indeed a more challenging one than classification.

1.6 Organization of this Document

The remainder of this thesis is structured as follows:

Chapter 2: We discuss previous work pertinent to the problem of analyzing sets of related time series data.

Chapter 3: We provide an introduction and brief review of the field of proteomics as it pertains to our work in Chapter 6. Readers not interested in this field can skip this chapter without much loss.

Chapter 4: We introduce the class of models called *Continuous Profile Models* (CPM) – a class of probabilistic generative models for alignment and normalization of sets of related time series data. We then present one specific instance of CPMs– the EM-based CPM (EM-CPM), as well as a method to train the EM-CPM. We explore use of the EM-CPM, mainly in the context of a liquid-chromatography mass spectrometry (LC-MS) data set, but also with a speech data set. Lastly, we briefly explore a multi-class version of the EM-CPM as applied to a NASA solenoid valve data set.

Chapter 5: We introduce a fully Bayesian instance of the class of CPM models, the *Hierarchical Bayesian Continuous Profile Model* (HB-CPM), as well as an associated (MCMC) algorithm for inference. We then explore use of this model on two data sets – a three-class liquid-chromatography-ultraviolet-diode array data from a study of the plant *Arabidopsis thaliana* and a two-class solenoid valve current data set.

Chapter 6: We present a simple technique for discovering differences between two classes of samples, which is used after data alignment. We apply this technique to LC-MS serum proteomic data without use of tandem mass spectrometry, gels, or labeling and test our technique on a controlled and realistic (spike-in, serum biomarker discovery) experiment which is therefore verifiable. This set-up allows us to assess different approaches to the alignment problem, by comparing precision-recall curves built from knowledge of the spike-in ground truth. We are thus able to contrast the performance of Dynamic Time Warping, the EM-CPM and the HB-CPM, demonstrating some advantages of CPMs.

Chapter 7: We wrap-up with a discussion of what we have accomplished and of future directions.

1.7 Notations and Conventions

Forms of the standard distributions not appearing in the text can be found in the first Appendix. Most background material is also found in various Appendixes rather than in the main manuscript so that readers already familiar with these topics can easily access the novel contributions of this thesis. Additionally, some technical details and derivations required for our models are relegated to various Appendixes in order to make for a clearer read.