Updating Markov models to integrate cross-sectional and longitudinal studies

Allan Tucker[a,*], Yuanxi Li[a], David Garway-Heath[b]

[a]*Department of Computer Science, Brunel University, UK*
[b]*Moorfields Eye Hospital and UCL Institute of Ophthalmology, University College London*

Abstract

Clinical trials are typically conducted over a population within a defined time period in order to illuminate certain characteristics of a health issue or disease process. Cross-sectional studies provide a snapshot of these disease processes over a large number of people but do not allow us to model the temporal nature of disease, which is essential for modelling detailed prognostic predictions. Longitudinal studies on the other hand, are used to explore how these processes develop over time in a number of people but can be expensive and time-consuming, and many studies only cover a relatively small window within the disease process. This paper explores the application of intelligent data analysis techniques for building reliable models of disease progression from both cross-sectional and longitudinal studies. The aim is to learn disease 'trajectories' from cross-sectional data by building realistic trajectories from healthy patients to those with advanced disease. We focus on exploring whether we can 'calibrate' models learnt from these trajectories with real longitudinal data using Baum-Welch re-estimation so that the

*Corresponding author
Email address: allan.tucker@brunel.ac.uk (Allan Tucker)

dynamic parameters reflect the true underlying processes more closely. We use Kullbaeck Liebler distance and Wilcoxon Rank metrics to assess how calibration improves the models to better reflect the underlying dynamics.

Key Words: Disease Progression, Cross-Sectional Studies, Markov Models

1. Introduction

Degenerative diseases such as cancer, Parkinson's disease, and glaucoma are characterised by a continuing deterioration to organs or tissues over time. This monotonic increase in severity of symptoms is not always straightforward however. The rate can vary in a single patient during the course of their disease so that sometimes rapid deterioration is observed and other times the symptoms of the sufferer may stabilise (or even improve - for example when medication is used). Interventions such as medication or surgery can make a huge difference to quality of life and slow the process of disease progression but they rarely change the long term prognosis. The characteristics of many degenerative diseases is therefore a general transition from healthy to early onset to advanced stages. Longitudinal studies [1] measure clinical variables from a number of people over time. Often, the results of multiple tests are recorded, generating Multivariate Time-Series (MTS) data. This is common for patients who have high risk indicators of disease where they are monitored regularly prior to diagnosis. For example, patients with high intra-ocular pressure are brought in to the clinic for visual field tests every six months as they are at high risk of developing glaucoma. The advantages of longitudinal data is that the temporal details of the disease progression

can be determined. However, the data is often limited in terms of the cohort size, due to the expensive nature of the studies. Cross-sectional studies record attributes (such as clinical test results and demographics) across a sample of the population, thus providing a snapshot of a particular process but without any measurement of progression of the process over time [2]. An advantage of cross sectional studies is that they capture the diversity of a sample of the population and therefore the degree of variation in the symptoms. The main disadvantage of such studies is that the progression of disease are inherently temporal in nature and the time dimension is not captured. For longitudinal analysis, the patients are usually already identified as being at risk and therefore, controls are usually not available and the early stages of the disease may have been missed. While many data integration techniques address representation heterogeneity where similar data is stored in many different forms, as is common in bioinformatics data [26], they do not attempt to combine variables from cross-section and longitudinal studies, which is what is the focus of this paper. A related area of research, known as panel analysis [21], involves trying to build models along both the temporal dimension and the population dimension from panel studies. Another line of research known as pooling has explored combining cross-sectional data with time-series data [22]. Fitting trends through data [23] is a common approach and is related in some ways to the idea of identifying a trajectory. Another related area of research is sequence reconstruction. This involves trying to find the best order for a particular set of data. Methods include the traveling-salesman-problem approach that aims to minimize the distance between each datum [24], and more recently, the use of PQ trees has been explored to en-

code partial orderings in order to account for uncertainty in the data due to elements such as noise [25]. Statistical process control [29, 28] has also been explored for modelling clinical data including data with unknown temporal ordering. Additionally, a resampling approach known as the Temporal Boot-Strap (TBS) [5] has been developed that aims to build multiple trajectories through cross sectional data in order to approximate genuine longitudinal data. These 'Pseudo Time-Series' (PTS) can then be used to build approximate temporal models for prediction. This approach has been extended in order to cluster important stages in disease progression using Hidden Markov Models (HMMs) [6]. However, the use of cross-sectional data alone will mean that no genuine timestamps have been used to infer the models and so they only capture an ordering without real temporal information.

In this paper, we explore how to minimise the expensive process of longitudinal data collection by taking models that are learnt from cross-sectional studies using pseudo temporal methods and 'calibrating' with limited longitudinal data. We do this calibration by using the Baum-Welch algorithm to update stochastic models learnt from pseudo time-series so that the dynamic parameters better reflect the underlying process. Essentially, we are integrating cross-sectional and longitudinal data to increase the temporal information and the diversity of data from a large population. Many data integration techniques address representation heterogeneity where similar data is stored in many different forms, as is common in bioinformatics data [7]. Meta Analysis, a popular approach [9], works by supplying a statistical framework for identifying significant results over a number of independent published studies, and calculating the significance of all of the studies when they are

brought together. However, it can be prone to publication bias where positive results are more likely to be published and therefore skew the statistics.

In the next section we formally describe the construction of pseudo time-series using the temporal bootstrap, the experimental set up for assessing the calibration of models with longitudinal data, and the clinical data from glaucoma patients that is used. In the results section, the added value of calibrating pseudo time-series models is demonstrated on simulated data and real clinical data. Finally a case study is explored using the longitudinal glaucoma data and a cross-sectional glaucoma study before conclusions are made.

2. Methods

2.1. Generating pseudo time-series

Let a dataset D be defined as a real valued matrix where m (rows) is the number of samples - here patients - and n (columns) is the number of variables - clinical test data. We define $D(i)$ as the ith row of matrix D. The vector $C = [c_1, c_2, \ldots, c_m]$ represents defined classes, where each $c_i \in \{0, 1\}$ corresponds to the sample i, $c_i = 0$ represents that sample i is a healthy case, and $c_i = 1$ represents that sample i is a diseased case. These classifications are based upon the diagnoses made by experts. We define a time-series as a real valued T (row) by n (column) matrix where each row corresponds to an observation measured over T time points. We say that if $T(i)$ was observed before $T(j)$ then $i < j$.

We define a set of pseudo time-series indices as $P = \{p_1, p_2, ...p_k\}$ where each p_i is a T length vector where $T > 0$. We define p_{ij} as the jth element

of p_i and each $p_{ij} \in \{1, ..., m\}$. We define the function $F(p_i) = [p_{i1}, \ldots, p_{iT}]$ as creating a T by n matrix where each row of $F(p_i) = D(p_{ij})$. A pseudo time-series can be constructed from each p_i using this operator. For example, if a pseudo time-series index vector $p_1 = [3, 7, 2]$ then $F(p_1)$ is a matrix where the first row is $D(3)$, the second row is $D(7)$ and the third row is $D(2)$. The corresponding class vector of each pseudo time-series generated by $F(p_i)$ is given by $G(p_i) = [C(p_{i1}), \ldots, C(p_{iT})]$.

To demonstrate this notation consider the following example:

Let the data matrix D be defined as:

$$D = \begin{vmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \\ d_{41} & d_{42} & d_{43} \end{vmatrix}, D_{ij} \in \Re.$$

Let the corresponding class vector be $C = [c_1, c_2, c_3, c_4]$. If $P = p_1, p_2$ where $p_1 = [1, 3, 1]$ and $p_2 = [2, 3, 1]$ then:

$$F(p_1) = \begin{vmatrix} d_{11} & d_{12} & d_{13} \\ d_{31} & d_{32} & d_{33} \\ d_{11} & d_{12} & d_{13} \end{vmatrix}, G(p_1) = [c_1, c_3, c_1].$$

and

$$F(p_2) = \begin{vmatrix} d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \\ d_{11} & d_{12} & d_{13} \end{vmatrix}, G(p_2) = [c_2, c_3, c_1].$$

Building pseudo time-series involves plotting trajectories through cross-sectional data based upon distances between each point using prior knowledge

of healthy and disease states. These trajectories can then be used to build temporal models such as Dynamic Bayesian Networks (DBNs) [10] and Hidden Markov Models (HMMs) to make forecasts [11]. The temporal bootstrap involves resampling data from a cross-sectional study and repeatedly building trajectories through the samples in order to build more robust time-series models. Each trajectory begins at a randomly selected datum from a healthy individual and ends at a random datum classified as diseased. The trajectory is determined by the shortest path of Euclidean distances between these two points. The data is first standardised to a mean μ of zero and a standard deviation σ of one as we found that this led to better HMM models. We use the Floyd-Warshall algorithm [12], a well established algorithm used to find the shortest path in a minimum spanning tree from the weighted graph. A full description of the algorithm to generate pseudo time-series is shown in Algorithm 1 below and appears in [5]. An example of pseudo time-series that have been generated from cross-sectional data are shown in Figure 1 below. Again, this was plotted on the first two components that were generated using multidimensional scaling.

2.2. The Experiments

We explore three sets of experiments that try to identify whether adding a small number of longitudinal data samples to models learnt from cross-sectional data (via the PTS approach outlined in Algorithm 1) improves them: i) One on simulated cross-sectional data whereby models are inferred using pseudo time-series and are compared to the original underlying time-series model. ii) Another on real data from Visual field tests where patients

Algorithm 1 PSEUDO TIME-SERIES ALGORITHM

Input: Cross section data D; class labels C, sample size T; number of pseudo time-series k

Standardise dataset D to $\mu = 0$ and $\sigma = 1$

for $i = 1$ to k **do**

Uniformly randomly sample (with replacement) T row indices from D to create d_i such that there is at least one healthy and one diseased class (in C) corresponding to any of the indices in d_i

Uniformly randomly select a row index from d_i, $start$, from where $1 \leq start \leq T$ and an endpoint, end, where $1 \leq end \leq T$ where $C(d_i, start)$ represents a healthy class and $C(d_i, end)$ represents a diseased class

Construct a TxT matrix, W_i, of Euclidean distances between each $D(d_{ia})$ and $D(d_{ib})$ for all combinations of indices in d_i

Calculate the minimum spanning tree over the matrix MST_i

Order d_i to create $d*_i$ based upon the shortest path between $D(d_i, start)$ and $D(d_i, end)$ given the tree MST_i using the FloydWarshall algorithm [21]

Add the ordered $d*_i$ to the set of pseudo time-series P

end for

return A set P of k pseudo time-series

Algorithm 2 CALIBRATING PSEUDO TIME-SERIES MODELS WITH LONGITUDINAL DATA

Input: Cross section data D; class labels C, Longitudinal Data E; sample size T; number of pseudo time-series k

Apply Algorithm 1 to generate a set P of k pseudo time-series

Run the Baum Welch algorithm until convergence to infer an Auto-Regressive Hidden Markov Model, H, from P

Update H using the Baum Welch algorithm with E for j iterations

return A calibrated Autoregressive HMM

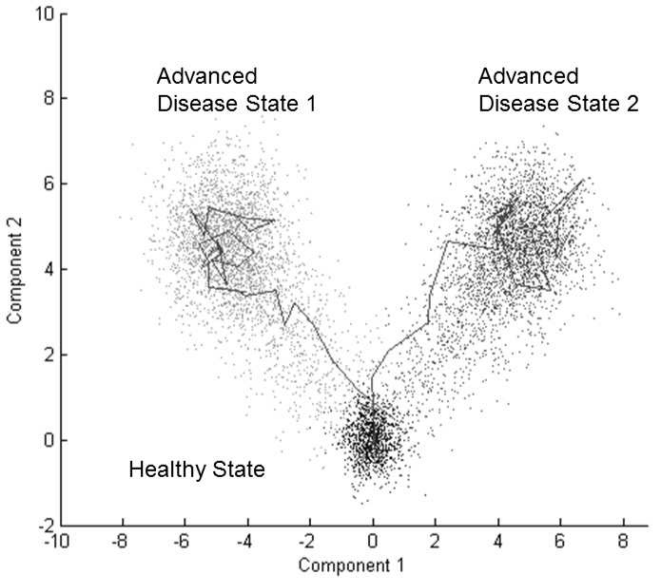

Figure 1: Example PTS generated from TBS on Simulated Data

who are at high-risk of developing glaucoma undertake a psychophysical test to identify damage to sectors of their vision. Here no true original model is known but a comparison can be made between single sampled-points of the time-series (to simulate a cross-section), and models learnt from the full time-series. iii) Finally, we explore integrating real cross-sectional clinical data with real longitudinal clinical data as a case study.

i) Simulated Data

Firstly, we explore the effect of *updating* models of cross-sectional data, built using PTS, with relatively small numbers of real time-series to see if the resulting models are improved. This involves the use of the Baum-Welch re-estimation algorithm applied to a prior HMM. This is outlined in Algorithm

2. Essentially we want to see if the limitations of pseudo time-series can be overcome (due to there being no time-element) by calibrating them with real time-series.

In detail, we generate time-series of length 30 from an AutoRegressive HMM (ARHMM) to mimic typical biomedical longitudinal data (MTS in Figure 2). We then randomly sample a single point from these series (CS DATA) to mimic the cross-sectional sampling of a population. We reserve 50 ARHMM time-series for the calibration (Reserved MTS). We start with 500 cross-sectional samples as this was found to be a suitably large size to generate good pseudo time-series and models in [5] and increment by 100 up to 1500 (the size of some increasingly large biomedical cross-sectional studies). We use the Kulbaeck-Leibler distance [13] to explore how close a model learnt from the cross-sectional data using the Temporal BootStrap (TBS) is to the original generating model. Finally, we use a number of the reserved time-series generated by the same ARHMM to update the pseudo time-series models (using Algorithm 2) and explore how close new calibrated models are to the original. Increments of 10 time-series were used as increments of this size seemed to involve significant changes in the KL distances. We also include how good the model is when learnt solely from the time-series used to calibrate the models.

ii) Clinical Test Data

We then apply a similar set of experiments with real clinical longitudinal data of visual fields from 91 patient time-series (91 MTS VF DATA in Figure 4).

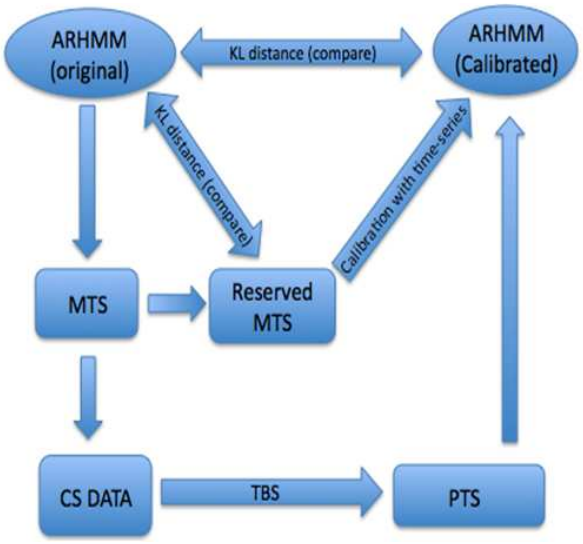

Figure 2: Simulated Data Experimental Framework

The longitudinal data is from a study of 23 ocular hypertensive patients (who eventually develop reproducible glaucomatous VF loss) from a longitudinal study at Moorfields Eye Hospital. A total of 255 patients with ocular hypertension (raised intraocular pressure, a major risk factor for glaucoma) volunteered to take part in a randomized placebo-controlled trial of treatment to prevent the onset of glaucoma [15]. Of these, a number developed reproducible VFs loss, as judged by the same classification algorithm, over a median period of six years. Subjects had several repeated clinical visits (approximately every six months). Each VF point maps to one of six Nerve Fibre Bundles (NFBs) where information from the retina leaves the eye and travels to the visual cortex [17] (see Figure 4). We average each VF points over their associated nerve fibre bundle to give 6 variables representing each spatial region. As a result, the data contains six NFB variables and one class variable.

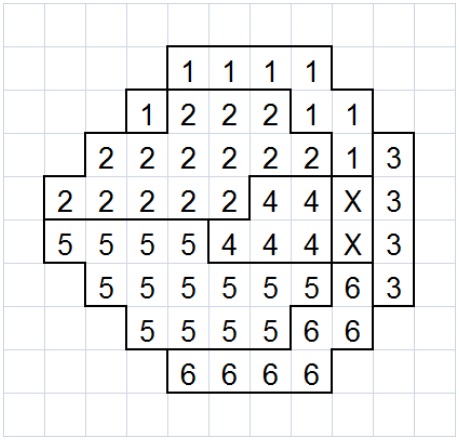

Figure 3: Spatial Distribution of the 6 Nerve Fibre Bundles over the Visual Field. X's Denote the Blindspot

We sample one VF test from each patient's time-series to generate a cross-sectional sample and generate pseudo time-series for learning a time-series model (PTS). We then compare this model as well as ones learnt from a combination of pseudo time-series and real time-series (Random 10/20 MTS) to see how quickly we can learn models that are close to the original. This is achieved by comparing these KL distances to the mean KL distance between 200 different ARHMMs learnt from the same original time-series (MEAN VARIANCE in Figure 4). In other words, if we can learn models from the sampled CS data that have similar KL distances to the general variation in learning a model from the full time-series, then we assume that the models are as close to one learnt from a full time-series.

iii) Clinical Data Integration

Finally, as a case study, we integrate the longitudinal data from the last experiment with real cross-sectional data in order to explore how the population-

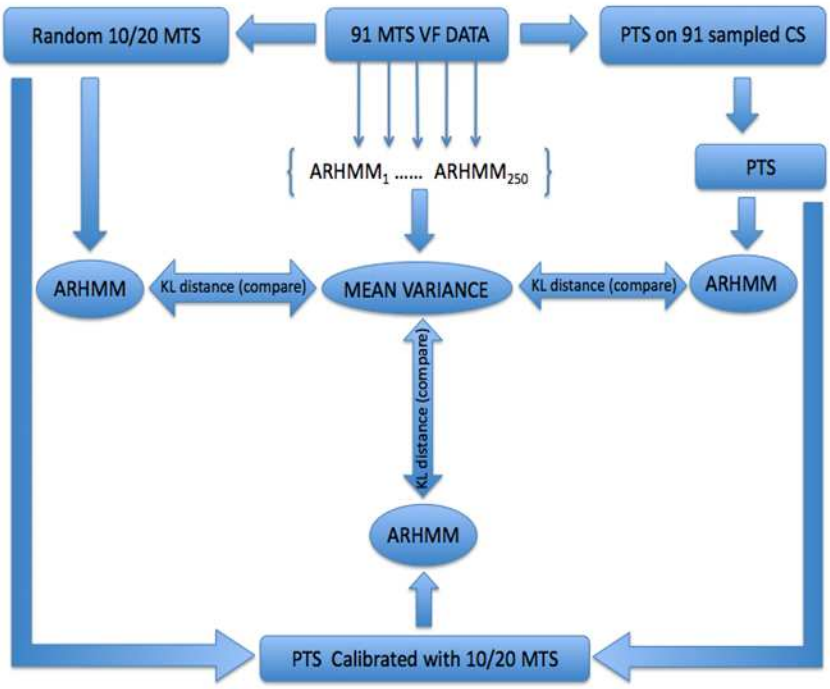

Figure 4: VF Data Experimental Framework

distribution information (from the cross-sectional data) and the dynamics of progression (from the longitudinal data) can be integrated successfully. The cross-sectional study consists of VF tests for 162 people, representing an expanded dataset that was used to evaluate the classification accuracy of an optic nerve head imaging device [18]. In brief, there were 84 healthy subjects and 78 patients with early glaucomatous VF loss. A full medical history was taken and a detailed ocular examination performed. Subjects underwent Humphrey VF testing with the 24-2 program [16]. The VF data for each subject are classified into one of two classes: healthy or glaucomatous based upon an established classification algorithm for the field test [16]. Again, the VF data is averaged into 6 NFBs as with the longitudinal study.

3. Results

3.1. Simulated Data Results

Figure 5 shows the results for learning PTS from cross-sectional samples of varying sizes and either not calibrating, or calibrating with 20 time-series, along with 95% confidence intervals.

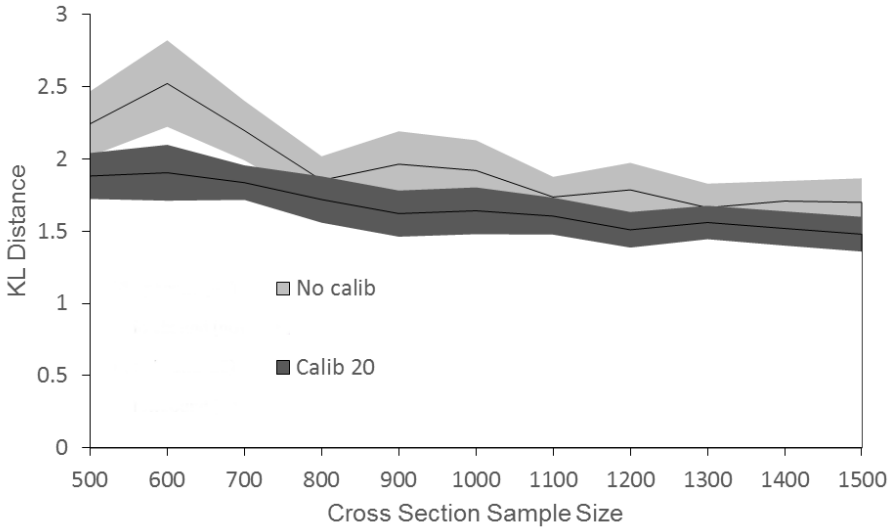

Figure 5: KL distance for varying cross-sectional study sample sizes with no calibration and with 20 longitudinal data samples for calibration.

The first obvious characteristic of these graphs is that calibrating does indeed improve the quality of the models with KL distances that are closer to the original generating ARHMM. This is not surprising seeing that there is no genuine 'time' in the PTS generated from the cross-sectional data. What is surprising, is that only a relatively small number of time-series are needed to improve these models, especially when there are lots of samples used from the cross-sectional data. This supports the results from previous

studies that the PTS does find good-but-not-perfect models (limited by the lack of real time-series) and that a small number of genuine time-series can calibrate these models. This offers hope that expensive longitudinal studies can be relatively small in size if combined with larger cross-sectional studies that capture the general trajectories and the variability of disease progression within a population. With calibration from 10 time-series, there is a steady decrease in KL distance as cross-sectional sample size increases where more and more reliable PTS are constructed. When the sample size is 1500 we see a KL distance mean of 1.70 ± 0.16. Note that when 10 time-series alone are used to learn the model we get a mean KL distance of 2.08 ± 0.26. This shows that the PTS generated from the cross-sectional data improves on models learnt from the time-series only by incorporating the variability within a larger population captured in the cross-sectional data. With calibration from 20 time-series we see a similar story, where increasing the cross-sectional sample size, build better PTS and results in models that are closer to the original. For 1500 in the cross-sectional sample we see a KL distance of 1.48 ± 0.12. Note that when 20 time-series alone are used to learn the model we get a mean KL distance of 1.78 ± 0.15. Again, it can be seen that the PTS improves on time-series alone but that the integration of both seems to generate the models that best reflect the underlying model. We now explore the statistical significance of the differences between these KL distances using the Wilcoxon Rank Comparison [19]. Figure 6 shows the Wilcoxon Rank statistic comparing the KL distance between different models learnt using the different approaches.

Wilcoxon Rank	cs500calib10	cs500calib20	cs1500nocalib	cs1500calib10	cs1500calib20	csfull30	csfull50
cs500nocalib	0.196	0.047	**0.001***	**0.001***	**0.001***	**0.001***	**0.001***
cs500calib10	-	0.455	0.062	0.036	**0.001***	**0.010***	**0.001***
cs500calib20	-	-	0.077	0.130	**0.001***	0.023	**0.001***
cs1500nocalib	-	-	-	0.947	0.119	0.395	0.064
cs1500calib10	-	-	-	-	0.052	0.277	0.047
cs1500calib20	-	-	-	-	-	0.395	0.728
csfull30	-	-	-	-	-	-	0.291
csfull50	-	-	-	-	-	-	-

Figure 6: Wilcoxon rank comparison between KL distances to original (significant p values are marked with an asterisk $p<0.01$)

An asterisk is used to denote significant p values (i.e. the models are significantly different). First of all notice that there are many significant values - implying that the difference between models learnt using the different approaches are significant. The most important statistics are those that show the models learnt with no calibration and only 500 cross-sectional data points are significantly different to most other models (row 1), but when 1500 cross-sectional data points are used the resulting model is much closer, only being significantly different to the model learnt from 50 full time-series (row 4). However, by calibrating these models we see improvement for 500 CS data points. For 1500 datapoints all models are not significantly different from the full 50 time-series, indicating that the PTS algorithm can find models that are not significantly different from a model inferred from full time-series data when sample size is high (though the uncalibrated model is significantly different at the 10% level - $p=0.064$). The model calibrated with 20 time-series (cs1500calib20) shows better improvement with a clearly insignificant difference between the models learnt from the full time-series ($p=0.728$).

3.2. Visual Field Data Results

We now explore the effect of calibrating PTS using the real Visual Field time-series data described earlier. As we have no knowledge of the true underlying model, we firstly compare the KL distance between models that are repeatedly learnt from the original 91 patient time-series in order to get an idea of general variance between models and to use this as a base-line. If we can generate models using PTS approaches with a KL distance that is not significantly greater than the general variance between different builds of the model on the full data, then it suggests that the PTS models are of a suitably similar quality to those learnt from the full time-series (note that variance in repeated model builds on full data could be due to small samples). We then calculate the KL distance between a model learnt from the sampled cross-section using the PTS approach and models learnt from the original 91 time-series. We then incrementally add a number of randomly selected real time-series to calibrate the PTS model to see if this improves the KL distance. We do this in two ways: simply concatenating the data (Concat), and also using the PTS as a prior which is updated with real time-series using the BW algorithm in Algorithm 2 (BW calibrated). Finally we calculate the KL distance between learning models using only the calibrating time-series to confirm that the PTS are indeed improving the resulting models. The experiments are repeated 100 times to derive confidence intervals on the KL distances. Figure 7 shows the results of these experiments.

Notice firstly that the KL distance between models that have been learnt on the full 91 time-series are in the region of 80-90 with a small confidence interval denoting a relatively small variance from one model learning to the

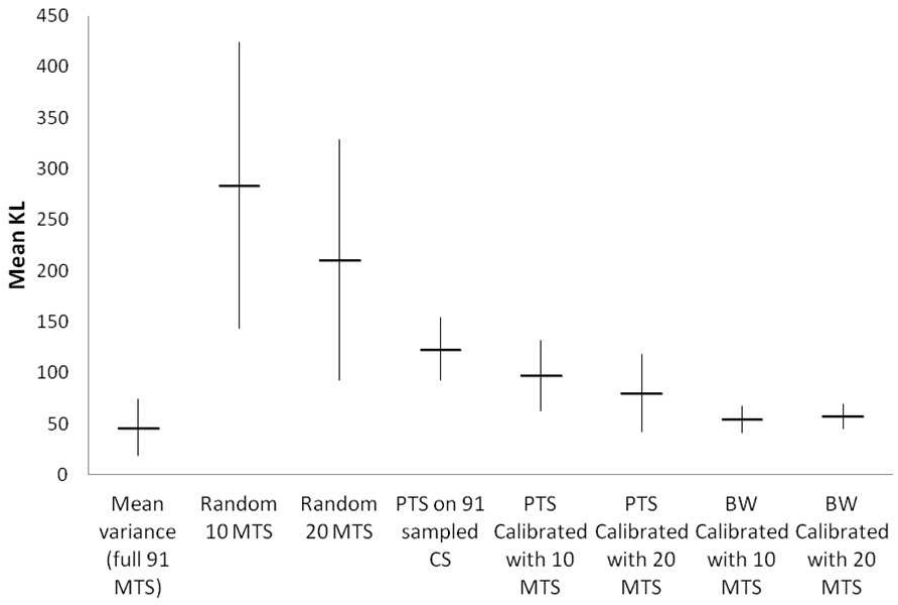

Figure 7: KL results for VF data with 95% confidence intervals.

Wilcoxon Rank	Rand 10	Rand 20	PTS	Concat (10)	Concat(20)	BW Cal(10)	BW Cal (20)
Mean variance (full 91 MTS)	**0.001***	**0.001***	**0.001***	**0.005***	0.011	0.255	0.500
Random 10 MTS	-	0.975	0.023	**0.002***	**0.001***	**0.001***	**0.001***
Random 20 MTS	-	-	0.042	0.014	0.010	**0.001***	**0.001***
PTS on 91 sampled CS	-	-	-	0.452	0.327	**0.001***	**0.001***
Concat with 10 MTS	-	-	-	-	0.773	**0.001***	**0.002***
Concat with 20 MTS	-	-	-	-	-	**0.002***	**0.006***
BW Calibrated with 10 MTS	-	-	-	-	-	-	0.881
BM Calibrated with 20 MTS	-	-	-	-	-	-	-

Figure 8: Wilcoxon rank siginificance (significant p values are marked with an asterisk $p<0.01$)

next. The models that are learnt from the sampled cross-section using the PTS approach are impressively close to the time-series models but distinctly higher in KL distance (likely to be because we are lacking real temporal information). When 10 and 20 real time-series are used to calibrate the model, however, we see further improvement in the KL distance resulting in models that are demonstrably closer to the models learnt from all 91 time-series. The updated models that go beyond simply concatenating data appear to per-

form the best with the lowest KL scores. Finally, models that are learnt from using the relatively small number of calibrating time-series only are clearly worse with much higher distance and large confidence intervals. Looking at the Wilcoxon Rank for significance as before, the important thing to notice in Figure 9 is that nearly all of the models are indeed significantly worse than the variation between models learnt on the full longitudinal dataset (significant differences are marked with an asterisk) except for the PTS model calibrated using the updating approach or concatenating with 20 real time-series. This shows that we can learn models that are as good as the natural variation between model building on the full longitudinal dataset by building PTS and calibrating with only 10 real longitudinal samples if we correctly balance the weighting of the cross-sectional PTS and real time-series. We can also see that many of the inferior models are similar in terms of their distances except for the very worst models (learnt from only 10 time-series) which are different from the superior models which are both PTS models that have been calibrated. To summarise, whilst the PTS approach alone does indeed learn very good models, by updating these models with a small number of real time-series we get models that are considerably closer to the models learnt using all the time-series data that is available. What is more, the Baum-Welch approach to updating improves upon a simply concatenation of data. Note that almost all models are significantly different from the general variance form learning the model from the full 91 time-series. The only models that are not significantly different at the 1% level are the PTS models updated with data using the Baum-Welch approach and the PTS model that is updated with 20 time-series by concatenation.

3.3. Case study: Integrating Visual Field Cross-section with Longitudinal Data

We now explore the use of a real cross-sectional dataset for building a PTS and calibrating with real longitudinal data. We apply the same process as outlined in Figure 4 but using the cross-sectional study discussed in methods, rather than sampling a cross-section from the longitudinal data. Calculating the KL distance to the full longitudinal data model is not applicable here as the cross-sectional study contains valuable information about healthy individuals and the early stages of disease that are not found in the longitudinal study. Therefore, a new gold-standard is required. For this we use the model learnt from the full cross-sectional data (using the pseudo time-series approach) but calibrated by the full longitudinal study. We then explore how few patient time-series are required to get close this standard. See Figure 9 for the results where it is clear that only a relatively small number of real MTS are required to get close to the gold-standard ($\geq$30).

We also explore the parameters of the different models: PTS, MTS and $Calib$. This includes the dynamic parameters for the underlying disease process and the static distributions for each nerve fibre bundle given either a healthy or a diseased diagnosis. These are shown in Figure 10 for the dynamic parameters (where we assume the longitudinal-only model as the gold-standard) and in Figure 11 for the static parameters (where we assume the larger cross-sectional-data-only model to be the gold-standard). Notice that the cross-sectional-only model (learnt using the pseudo time-series approach - PTS) has learnt distributions for the dynamic parameters that are surprisingly close to the gold-standard (MTS) model, but that the probability of

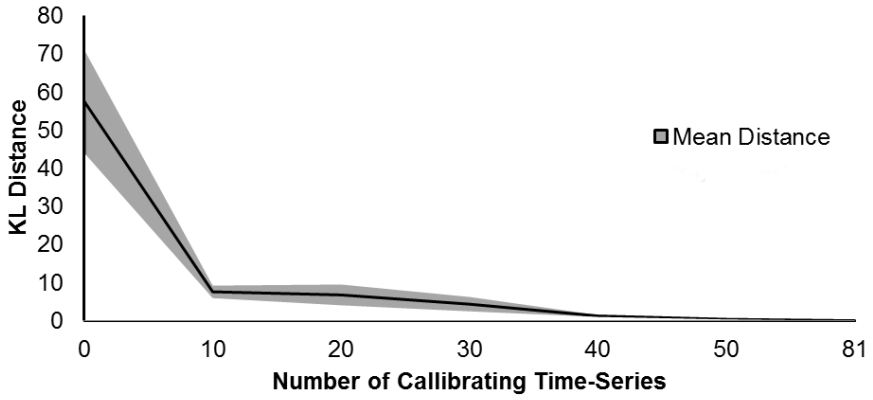

Figure 9: KL Distance to the model learnt from the real CS Data and calibrated with the full real longitudinal data, for differing numbers of calibrating MTS - including 95% confidence intervals

switching from healthy to glaucomatous is too high. The calibration of this model with time-series ($Calib$) improves this distribution considerably with a closer match to the gold-standard. The static parameters for each NFB show that the model learnt from the full longitudinal study (MTS) is sometimes very different from the full cross-sectional model (PTS) which is considered the gold-standard for distributions over the different NFBs (distributions for NFBs 1 and 2 in particular are very different with the gold-standard PTS being biassed to low VF sensitivity, but the MTS biassed to higher VF sensitivities). The calibrated model demonstrates a set of distributions that are generally closer to the gold-standard. For example, NFBs 3, 4 and 6 are much closer than the uncalibrated MTS model. NFB 5 shows a slight improvement in the distribution when the MTS is calibrated. Looking at the spatial layout of these NFBs in Figure 3 it seems that the distributions that are not easily learnt from the MTS data are those in the upper hemisphere (NFBs 1 and 2). This is interesting in that it is often these where the early

signs are glaucoma are first detected.

In summary, the calibrated model better represents the gold standard for the dynamic parameters (learnt from the longitudinal data) and the static parameters for each NFB (learnt from the cross-sectional data), though some NFBs show better improvement than others.

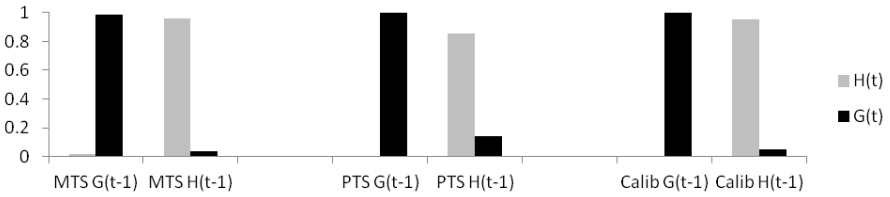

Figure 10: Dynamic Parameters for hidden variable where we consider the MTS model learnt from the longitudinal data as the gold standard

4. Conclusions

In this paper we have explored to what degree pseudo time-series, learnt from building trajectories through a cross-sectional study, can be 'calibrated' by a relatively small number of real time-series data form a clinical longitudinal study. The aim is to gain the advantage of both types of study - the population diversity of symptoms at all stages of a disease process from cross-sectional data; and the inherently temporal information of a disease process from longitudinal data. We have demonstrated that a relatively small number of disease time-series can dramatically improve the quality of disease model if the pseudo time-series has been constructed from a large enough cross-sectional sample. This has been shown to be the case for simulated data based upon a probabilistic model and real-world clinical data where the resultant models are not significantly different to models learnt from large

longitudinal studies. The approach is best suited to large cross-sectional studies though we have shown that only a small number of longitudinal samples are necessary to achieve improvement. Data has to be standardised prior to building the trajectories and the distance metrics used in this paper to construct them are based upon real valued data. In order to deal with binary / discrete data which is also common in medical contexts, other metrics could be explored such as the Jaccard index, kapp, and adjusted rand.

Sometimes it will be important to place constraints on the trajectories generated by the pseudo time-series. For some datasets, our method could potentially build impossible trajectories: For example, a trajectory could be built that contains different patients with different antibodies. As these should not change in an individual over time, it makes the trajectory unrealistic. Different mechanisms to constrain these trajectories will be important. One way to do this could exploit more detailed clinical evaluation rather than the simplistic labelling of healthy and post-diagnosis. For example, sometimes severity of stages in a disease progression are available and these can be used to guide the trajectory construction.

Pseudo time-series naturally model multiple endpoint analysis which is an important topic in modelling disease progression [20]. Future work will explore the explicit understanding of these in terms of identifying subcategories of disease (which they may well represent) and which we have already started to explore [6]. We are also interested in exploring latent variables in the context of discovered trajectories in order to identify subclasses similar to [27] who use them in the context of extended mixed models.

References

[1] Albert, PS. Longitudinal data analysis (repeated measures) in clinical trials. Statistics in medicine, 18(13):1707-1732, 1999.

[2] Mann, CJ. Observational research methods. research design ii: cohort, cross sectional, and case-control studies. Emergency Medicine Journal, 20(1):54-60, 2003.

[3] Frank A and Asuncion, A. UCI machine learning repository. Irvine: University of California, school of information and computer science. Available at: http://archive.ics.uci.edu/ml , 2010, Last accessed 17th Dec 2013.

[4] Seber, GAF. In Multivariate Observations. John Wiley and Sons, Hoboken, NJ, 1984.

[5] Tucker, A. and Garway-Heath, D. The pseudo temporal bootstrap for predicting glaucoma from cross-sectional visual field data, IEEE Trans IT Biomed, 14 (1) (2010), pp. 79-85

[6] Li, Y. and Swift, S. and Tucker, A., Modelling and analysing the dynamics of disease progression from cross-sectional studies, Journal of Biomedical Informatics 46 (2) : 266- 274, 2013

[7] Shen, R. et al., Integrative Subtype Discovery in Glioblastoma Using iCluster, PLOS ONE, 7 (4), e35236, 2012.

[8] Inmon, WH. Building the Data Warehouse. John Wiley and Sons, 2nd edition, 1996.

[9] Steele, E. and Tucker, A. Consensus and meta-analysis regulatory networks for combining multiple microarray gene expression datasets, Journal of Biomedical Informatics 41 (6) : 914- 926, 2008

[10] Murphy, K. Dynamic Bayesian Networks: Representation, Inference and Learning, PhD Thesis, University of Califronia, Berkeley, 2002.

[11] Rabiner, LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 257-286, 1989.

[12] Floyd. RW. Algorithm 97: shortest path. Communications of the ACM, 5(6):345, 1962.

[13] Kasza, J. and Solomon. PJ. A comparison of score-based methods for estimating Bayesian networks using the Kullback Leibler divergence. arXiv:1009.1463v2 [stat.ME]. In press for Communications in Statistics: Theory and Methods, 2013.

[14] Efron, B. Tibshirani, R. An introduction to the bootstrap (monographs on statistics and applied probability) CRC Press, Boca Raton, FL (1993)

[15] Kamal, D. Garway-Heath, D.F. Ruben, S. OSullivan, F. Bunce,C. Viswanathan, A. Franks, W. and Hitchings, R. Results of the betaxolol versus placebo treatment trial in ocular hypertension, Graefes Arch. Clin. Exp. Ophthalmol., Vol. 241, pp. 196203, 2003.

[16] AGIS, Advanced glaucoma intervention study. 2. visual field test scoring and reliability, Ophthalmology, Vol. 101, no. 8, pp. 14451455, 1994.

[17] Garway-Heath, D.F. Fitzke, F. and Hitchings, R.A. Mapping the visual field to the optic disc, Ophthalmology, vol. 107, pp. 18091815, 2000.

[18] Wollstein, G. Garway-Heath, D.F. and Hitchings, R.A. Identification of early glaucoma cases with the scanning laser ophthalmoscope, Ophthalmology, vol. 105, pp. 15571563, 1998

[19] Bauer, DF. Constructing confidence sets using rank statistics. Journal of the American Statistical Association 67, 687-690, 1972.

[20] Pocock, J., Stuart, L., Geller N., and Anastasios, AT. The Analysis of Multiple Endpoints in clinical trials. Biometrics, 43:487-498, 1987.

[21] D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh, Panel Surveys. New York: Wiley, 1989.

[22] R. Fanfani, Pooling time-series and cross-section data: A review, Eur.Rev. Agric. Econ. , vol. 2, no. 1, pp. 6385, 1974.

[23] M. Bianchi, M. Boyle, and D. Hollingsworth, A comparison of methods for trend estimation, Appl. Econ. Lett., vol. 6, no. 2, pp. 103109, 1999.

[24] S. S. Skiena, Traveling Salesman Problem. Berlin, Germany: Springer-Verlag, 1997, pp. 319322.

[25] M. Magwene, P. Lizardi, and J. Kim, Reconstructing the temporal ordering of biological samples using microarray data, Bioinformatics, vol. 19, no. 7, pp. 842850, 2003.

[26] R. Alfieri, I. Merelli, E. Mosca, and L. Milanesi. A data integration approach for cell cycle analysis oriented to model simulation in systems biology. BMC Systems Biology, 1:135, 2007.

[27] C. Proust-Lima, V. Philipps, B. Liquet, Estimation of extended mixed models using latent classes and latent processes: the R package lcmm, Cornell University Library arXiv:1503.00890v2 [stat.CO]

[28] L. Minne, S. Eslami, N. de Keizer, E. de Jonge, S.E. de Rooij, A. Abu-Hanna, Statistical process control for validating a classification tree model for predicting mortality A novel approach towards temporal validation, Journal of Biomedical Informatics, 45(1), 2012, Pp. 3744

[29] A.J Poots and T. Woodcock, Statistical process control for data without inherent order, BMC Medical Informatics and Decision Making, 201212:86, DOI: 10.1186/1472-6947-12-86

[30] The Advanced Glaucoma Intervention Study Investigators. Advanced Glaucoma Intervention Study. 2. Visual field test scoring and reliability. Ophthalmology 1994; 101(8):144555.

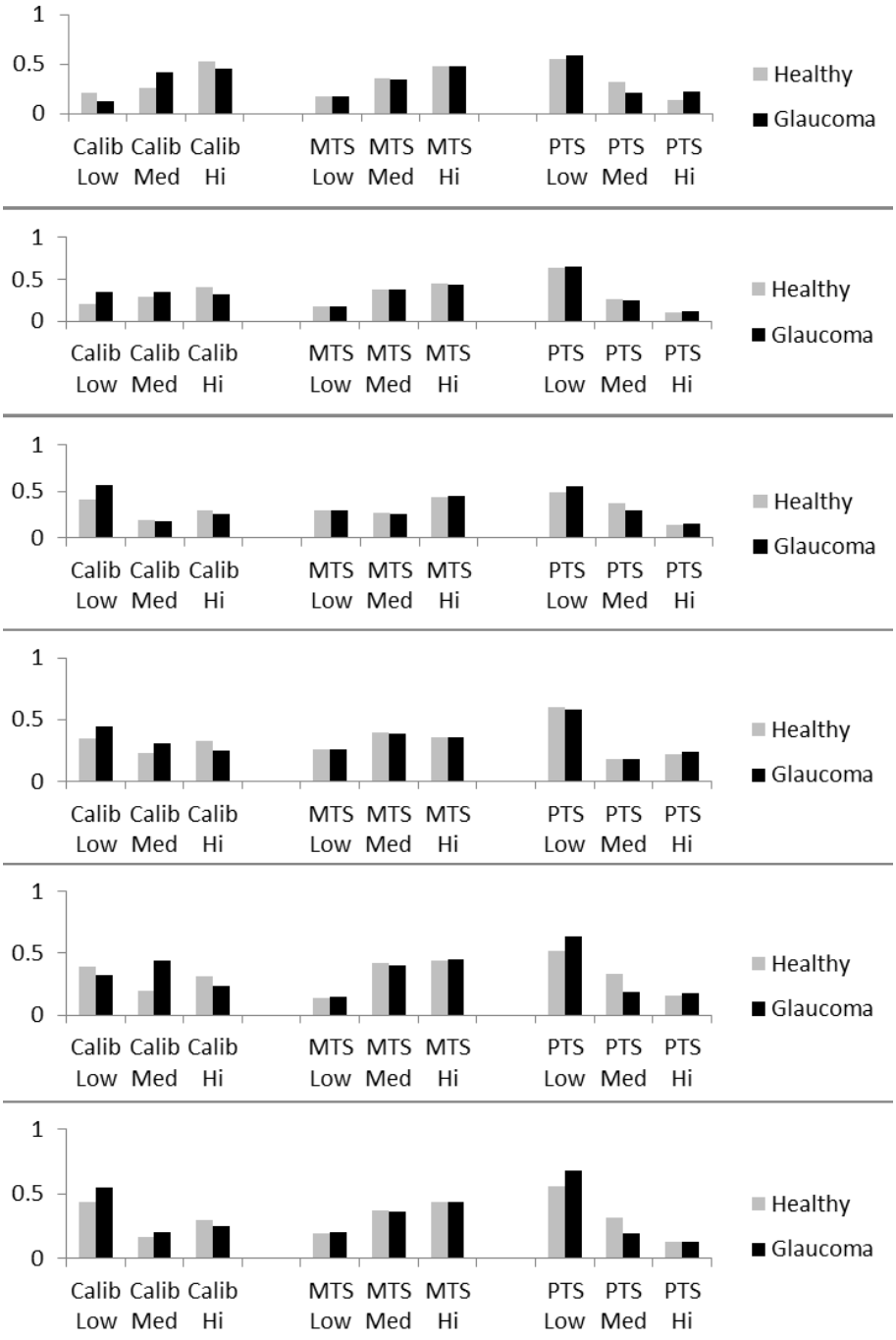

Figure 11: Static parameters for the 6 NFBs where we consider the PTS model learnt from the large CS study as the gold standard. These variables take on three discrete states: Low, Medium, and High sensitivity