Updating stochastic networks to integrate cross-sectional and longitudinal studies

Allan Tucker and Yuanxi Li
allan.tucker@brunel.ac.uk

Department of Computer Science, Brunel University, UK

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. These 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.

Key Words: Disease Progression, Cross-Sectional Studies, Stochastic Networks

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 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 as is explored in [3]. Previously, a resampling approach known as the Temporal BootStrap (TBS) [6] 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) [7]. 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 updating with limited longitudinal data. We do this by using the Baum-Welch algorithm to update stochastic models learnt from pseudo time-series. 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 different forms, as is common in bioinformatics data [8].

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 $C = [c_1, c_2, \ldots, cm]$ 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 $P = \{p1, p2, ...pk\}$ where each pi 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 pi 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 $G(pi) = [C(p_{i1}), \ldots, C(P_{iT})]$.

To summarise we have defined a set of k pseudo time-series with their associated class labels, sampled from the cross sectional data D indexed by the elements of pi. 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 weighted graph. A full description of the algorithm to generate pseudo time-series appears in [6] and an example of pseudo time-series that have been generated from cross-sectional data are shown in Figure 2. Again, this was plotted on the first two components generated using multidimensional scaling.

2.2 The Experiments

We explore two set of experiments that both try to identify whether adding a small number of longitudinal data samples to models learnt from cross-sectional data (via the PTS approach) improves them: i) One on simulated data whereby discovered models are compared to some original time-series model from where the data has been sampled. ii) Another on real data from visual field tests where patients who are at high-risk of developing glaucoma undertake a psychophysical test to identify damage to sectors of their vision. Apart from the clinical test data itself and the clinical diagnosis (glaucomatous or not), no demographic data is included. Here no true original time-series model is known but a comparison can be made to models learnt on the time-series and on sampled cross-sections of the time-series. 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. Essentially we want to see if the limitations of pseudo time-series can be overcome (due to there being

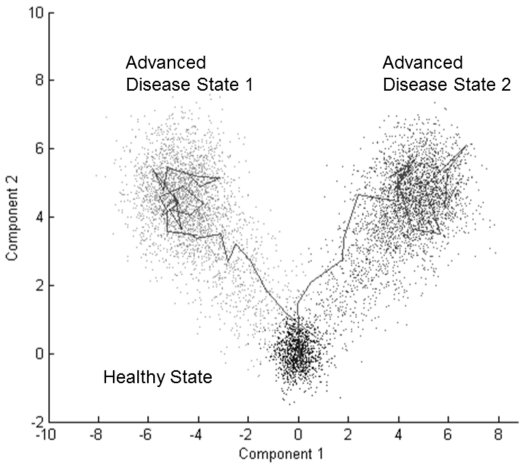


Fig. 1. Example PTS generated from TBS on Simulated Data

no time-element) by calibrating them with real time-series. We generate time-series of length 30 from an AutoRegressive HMM (ARHMM) to mimic typical biomedical longitudinal data (MTS in Figure 3). We then randomly sample a single point from these series (CS DATA) to mimic the cross-sectional sampling of a population but reserve 50 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 [6] 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. Experiments involve bootstrapping the data [14] We then add a number of the reserved original time-series generated by the same ARHMM to the pseudo time-series and explore how close new calibrated models are to the original. Increments of 10 time-series were used as these seem to differentiate between the KL distances significantly. We also include how good the model is when learnt solely from the time-series used to calibrate the models. We then apply a similar set of experiments with real VF longitudinal with 91 patient time-series (91 MTS VF DATA in Figure 4). We sample one VF test from each patient's time-series to generate a cross-sectional sample and generate PTS data for learning models from (PTS). We then compare this model as well as ones learnt from a combination of PTS 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 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-

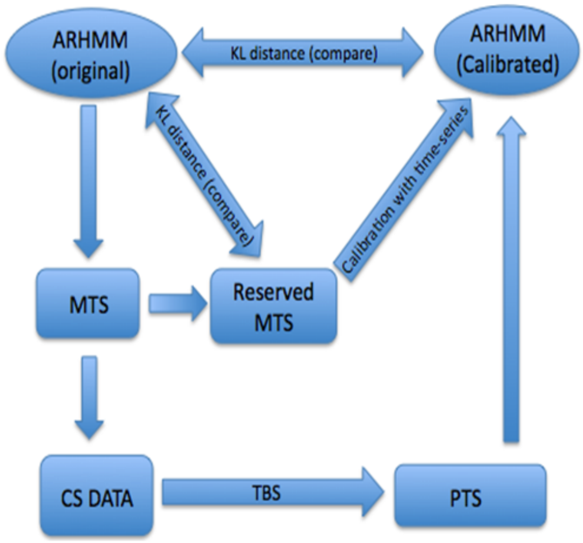


Fig. 2. Simulated Data Experimental Framework

series, then we assume that the models are as close to one learnt from a full time-series.

3 Results

3.1 Simulated Data Results

Figure 5 shows the results for all experiments, learning PTS from cross-sectional samples of varying sizes and either not calibrating, calibrating with 10 time-series, or calibrating with 20 time-series. 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

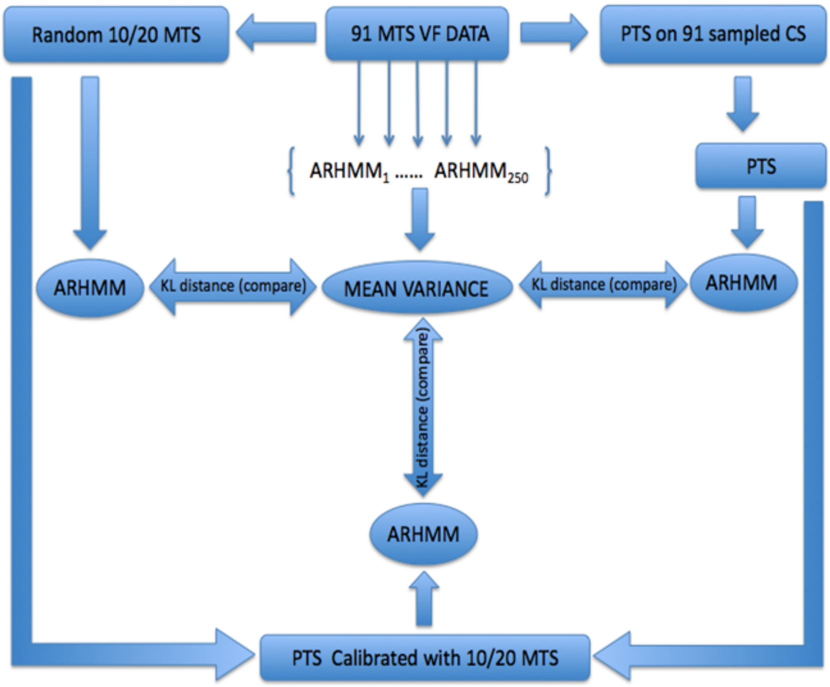


Fig. 3. VF Data Experimental Framework

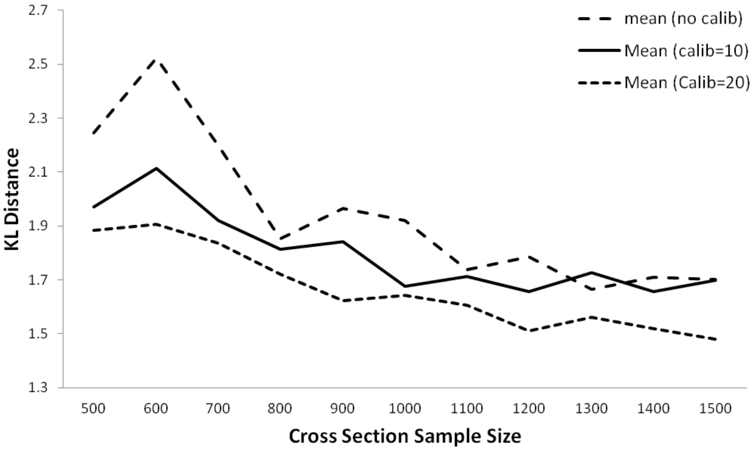
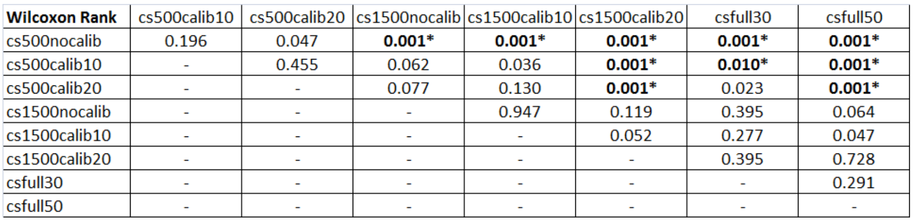


Fig. 4. KL distance for varying cross-sectional study sample sizes with increasing number of longitudinal data for calibration.

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 [15]. Table I shows the Wilcoxon Rank statistic comparing the KL distance between different models learnt using the different approaches. An asterisk is used denote significant p values. First of all notice that there are many significant values - indicating 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 is significantly different to most other models (row 1), but when 1500 cross-sectional data points are used the resulting model is much closer, being most different (though not quite significantly) to the model learnt from 50 full time-series (row 4). By calibrating these models we see a little improvement for 500 CS data points. However, for 1500 datapoints calibrated with 20 time-series, there is no significant difference between the models learnt from the full time-series with much higher p-values.



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	-	-	-	-	-	-	-

Fig. 5. Wilcoxon rank comparison between KL distances to original MTS model (significant p values marked with an asterisk P<0.01)

3.2 Visual Field Data Results

We now explore the effect of calibrating PTS using the real Visual Field time-series data described earlier (Figure 6). 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. Essentially, 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 we can be confident that the PTS models are of a suitably similar quality to those learnt from the full time-series. 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 (PTS calibrated), and also using the PTS as a prior which is updated with real time-series - Baum Welch (BW) calibrated Bayes. 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

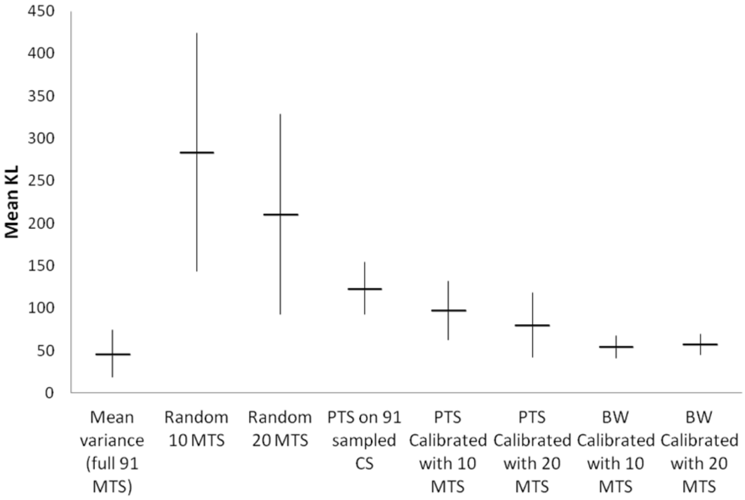
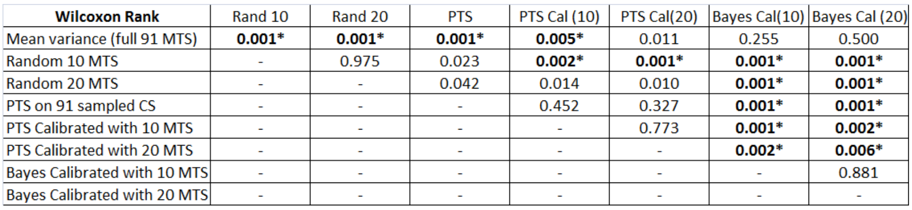


Fig. 6. KL results for VF data with confidence intervals.



Wilcoxon Rank	Rand 10	Rand 20	PTS	PTS Cal (10)	PTS Cal(20)	Bayes Cal(10)	Bayes 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***
PTS Calibrated with 10 MTS	-	-	-	-	0.773	**0.001***	**0.002***
PTS Calibrated with 20 MTS	-	-	-	-	-	**0.002***	**0.006***
Bayes Calibrated with 10 MTS	-	-	-	-	-	-	0.881
Bayes Calibrated with 20 MTS	-	-	-	-	-	-	-

Fig. 7. Wilcoxon rank for VF data (significant p values marked with an asterisk P<0.01)

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 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 perform 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 Table II 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.

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. Future work will involve exploring the Baum-Welch updating approach to integrating the longitudinal and cross-sectional data on more datasets. Pseudo time-series naturally model multiple endpoint analysis which is an important topic in modelling disease progression [16]. Future work will explore the explicit understanding of these in terms of identifying subcategories of disease which we have already started to explore [7].

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. Siddiqui, ZF. et al. Predicting the post-treatment recovery of patients suffering from traumatic brain injury (TBI), Brain Informatics 2:33-44, 2015.
4. 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.
5. Seber, GAF. In Multivariate Observations. John Wiley and Sons, Hoboken, NJ, 1984.
6. 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
7. 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
8. Shen, R. et al., Integrative Subtype Discovery in Glioblastoma Using iCluster, PLOS ONE, 7 (4), e35236, 2012.
9. Inmon, WH. Building the Data Warehouse. John Wiley and Sons, 2nd edition, 1996.
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. Bauer, DF. Constructing confidence sets using rank statistics. Journal of the American Statistical Association 67, 687-690, 1972.
16. Pocock, J., Stuart, L., Geller N., and Anastasios, AT. The Analysis of Multiple Endpoints in clinical trials. Biometrics, 43:487-498, 1987.