Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Evaluating reliability of hidden Markov models that describe the lifting patterns of chronic lower back pain patients and controls

Jill C. Slaboda, J. Robert Boston, *Member, IEEE* and Thomas E. Rudy

***Abstract*— Two hidden Markov models (HMMs) were designed to identify sub-groups of chronic lower back pain (CLBP) subjects based on time series of lifting parameters obtained during a repetitive lifting task. Two simulation studies were conducted to determine the reliability of this approach, using data from the repetitive lifting study. The first simulation verifies that control and CLBP HMMs based on these data can reliably identify sequences that were generated from that model. The second simulation determines whether the HMMs can reliably identify sequences that are intentionally misclassified (CLBP lifting sequences included in the control group and visa versa). The kappa statistic is used to quantify reliability. The simulation results show that the HMMs provide a reliable technique to analyze time series of lifting patterns and can be used to identify misclassified subjects as a sub-group.**

I. Introduction

Clinical studies that investigate the impact of chronic lower back pain (CLBP) on physical functioning commonly involve comparisons of CLBP and control groups, where both groups are treated as homogeneous. In fact these groups, particularly the CLBP group, are probably not homogeneous. This paper describes a method that uses hidden Markov models (HMMs) to identify groups within a CLBP population based on individual time series data. The method is being developed to characterize subjects in a study of the impact of CLBP on physical functioning.

Psychosocial research has identified groups of CLBP patients based on their responses to self-reported measures suggesting that CLBP patients are heterogeneous. The West Haven-Yale Multidimensional Pain Inventory (MPI), which assesses pain-relevant psychosocial aspects, responses of the significant other to the patient's pain and frequency of common activities indicated three distinct groupings of CLBP patients based on the patient's responses [1]. These groups were: Dysfunctional group, characterized by higher levels of pain, life interferences, emotional distress and functional limitation; Interpersonally distressed group, characterized by lower levels of social and personal support; and Adaptive copers group, characterized by lower levels of pain, functional limitation, and emotion distress. The MPI is a commonly used measure in the chronic pain literature and has been shown to be reliable and valid [2]. The CLBP sub-groups identified with the MPI demonstrate the heterogeneity of CLBP patients. Despite this evidence, differences in motion among CLBP subjects have received little attention in biomechanical research.

Temporal changes in lifting parameters have been found to differ between CLBP subjects and control subjects. Rudy et al. examined changes in lifting parameters over the duration of a repetitive lifting task by separating task time into three phases of early, middle and late [3]. CLBP subjects made adaptations to lifting parameters in the early to middle phases of the task and control subjects made adaptations throughout all task phases. Similar results were found in Slaboda et al., who analyzed jerk at the shoulder of control and CLBP subjects performing a repetitive lifting task [4]. CLBP subjects were found to perform lifts with lower jerk values than controls, and over task time, CLBP subjects increased jerk from the early to middle phases of the task while control subjects increased jerk throughout all phases of the task. Vasko et al used HMMs to describe the time series of lifting patterns of CLBP subjects and control subjects during a repetitive lifting task [5]. The HMM describing control subjects' motion was found to have a topology that contained more transitions than the HMM topology that described lifting motion of CLBP subjects.

Since CLBP subjects have been found to demonstrate different temporal motion patterns than control subjects during a repetitive lifting task and HMMs provide a useful technique to represent these time series data, we propose to use HMMs to identify sub-groups in the CLBP patient and control populations. Specifically, we propose an approach to identify a sub-group of CLBP subjects whose lifting data is more similar to control subjects than to other CLBP subjects (high performing CLBP subjects) and a sub-group of control subjects whose lifting data is more similar to CLBP subjects than to other control subjects (low performing controls). Two HMMs, one for each group, were trained using a jackknife method [6]. This method excludes a lifting sequence of a single subject and trains the two HMMs with the remaining control subject data and CLBP subject data. The excluded lifting sequence is tested with each model to determine the likelihood that the model produced the sequence. The subject is then classified to the group whose model provides the greatest likelihood probability.

This paper describes two simulation studies to determine the reliability of this approach, using data from the repetitive lifting study. The first simulation verifies that the CLBP

Manuscript received April 24, 2006. This work was supported by Research Grant 1R01 AG18299 from the National Institute on Aging, National Institute of Health, Bethesda, MD 20892, USA.

J. C. Slaboda is with Bioengineering Department, University of Pittsburgh, PA 15261, USA. (phone: 412-655-8057; fax: 412-665-8067; e-mail: jcsst46@pitt.edu)

J. R. Boston is with Electrical Engineering and Bioengineering Departments, University of Pittsburgh, PA 15261, USA. (e-mail: boston@engr.pitt.edu)

T. E. Rudy is with Departments of Anesthesiology, Psychiatry and Biostatistics, Pain Evaluation and Treatment Institute, University of Pittsburgh, PA 15206, USA. (e-mail: RudyTE@anes.upmc.edu)

HMM and control HMM based on these data can reliably identify sequences that were generated from that model. The second simulation determines whether the HMMs can reliably identify sequences that are intentionally misclassified (CLBP lifting sequences included in the control group and visa versa). The kappa statistic is used to quantify reliability.

II. METHODS

The HMMs were designed to describe the time series of lifting patterns that were performed by control and CLBP subjects during a repetitive lifting task. A total of 105 subjects (51 pain-free control subjects and 54 CLBP subjects, aged 21-65 years) completed the repetitive lifting task after giving informed consent as approved by the University of Pittsburgh Biomedical Institutional Review Board. The task required subjects to repeatedly lift the handle of a work simulator (Baltimore Therapeutic Equipment Company, Baltimore, MD, USA) from 13 inches above the floor to waist height. The work simulator provided resistance during the up–phase of the lift. The resistive load was set to 40% of the subject's static strength. Subjects performed the lifting task for 20 minutes with a 15-second rest period between each lift. Reflective markers were placed on the subjects' ankle, knee, hip, shoulder and tracked at 30 frames per second (Motion Analysis Corporation, Santa Rosa, CA, USA). Several lifting parameters (knee midpoint, hip midpoint, starting hip angle, starting knee angle, root-mean-squared jerk, time when maximum jerk occurred, hip-knee midpoint difference and lift duration) were calculated from the marker displacements for each lift performed during the task, resulting in a time series of lifting parameters for each subject. The time series data were described by HMMs.

In order to design HMMs with simple structure, the HMMs were constructed from the results of a data reduction procedure. This procedure used factor analysis to reduce the redundant parameters into four factors, and k-means cluster analysis to assign each lift to one of five clusters based on the four factor scores of the lift. The number of clusters was determined based on statistical techniques suggested by Mulligan et al [7]. The clusters defined the observation sequence of the subjects and output tokens of the HMMs. Using the results of the cluster analysis, two HMMs (one describing lifting patterns of controls and one describing lifting patterns of CLBP subjects) were designed and tested in the simulation studies. Three states were used, based on the results from Vasko et al., who determined that a three-state HMM best described temporal lifting patterns of control and CLBP subjects during a lifting task [5].

To determine the initial estimates of the HMM parameters, the four factor scores of each lift were separated by group and k-means clustered into three states. The estimates of the token probabilities were determined from a histogram of the five tokens within each of the states for all CLBP subjects and controls separately. The estimates of the transition probabilities were calculated as the total number of transitions of all the subjects within the group from state i to state j divided by the sum of the transitions in state i.

The HMMs were trained using the Baum Welch training algorithm, with Rabiner's multiple sequence observation method [8]. The control HMM was trained with the control subject data and the CLBP HMM was trained with the CLBP subject data. The Baum Welch training algorithm was set to 500 iterations and stopped when the probabilities converged to a difference of 10^{-6}. Both of the trained HMMs were fully connected ergodic models as shown in Figure 1, with parameters shown in Table 1.

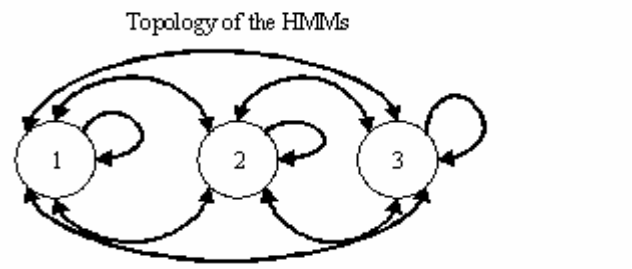

Figure 1: Topology of the control HMM and CLBP HMM

Table 1: Parameters of control HMM and CLBP HMM

Control HMM					
Control transition probabilities					
	Transition to state 1		Transition to state 2		Transition to state 3
In state 1	0.937		0.016		0.047
In state 2	0.013		0.982		0.006
In state 3	0.056		0.015		0.929
Control Token Probability					
	Token 1	Token 2	Token 3	Token 4	Token 5
State 1	0.142	0.082	0.037	0.036	0.703
State 2	0.002	0.029	0.863	0.04	0.067
State 3	0.012	0.537	0.050	0.401	0
CLBP HMM					
CLBP transition probabilities					
	Transition To state 1		Transition to state 2		Transition To state 3
In state 1	0.957		0.022		0.022
In state 2	0.015		0.976		0.009
In state 3	0.008		0.003		0.989
CLBP Token Probability					
	Token 1	Token 2	Token 3	Token 4	Token 5
State 1	0.932	0.008	0.016	0.019	0.025
State 2	0.053	0	0.049	0.841	0.057
State 3	0	0.407	0.105	0.006	0.481

Simulated lifting sequences were generated from the trained token and transition probabilities using a random number generator in MatLab (MathWorks Incorporated, Natick, Massachusetts, USA). A kappa statistic was used to evaluate reliability in both simulation studies because this statistic assesses agreement after adjusting for chance agreement [9]. Kappa was calculated as the observed

probability of agreement subtracted from the probability of chance agreement divided by one minus the probability of chance agreement. Values of kappa that are > 0.8 indicate excellent reliability, between 0.8 and 0.6 indicate substantial reliability, between 0.6 and 0.4 indicate moderate reliability, between 0.4 and 0.2 indicate fair reliability and kappa statistics lower than 0.2 indicates slight to no reliability. In this study, kappa ≥ 0.6 was considered to be reliable.

A. First Simulation

In the first simulation study, simulated lifting sequences were generated and tested against both models. The simulation was run three times so that the starting state could be set to a different state in each trial. The starting state was varied between trials but not within the trial (i.e. trial 1 tested sequences that all started in state 1 etc.) To evaluate the influence of the number of sequences and the length of the sequences on reliability in each trial, the number of simulated sequences in each group was varied in increments of 10 from 20 sequences (10 simulated from control HMM and 10 simulated from the CLBP HMM) to 120 sequences (60 from control HMM and 60 from the CLBP HMM) and the number of lifts in the sequences was varied from 6 to 80 lifts in each of the three trials. The maximum sample size (120) and range of lifts in each sequence were selected to match the data from the clinical study that will ultimately be tested. Each simulated sequence was tested against both HMMs and classified to the model with the greater likelihood probability that the sequence was observed given the model parameters. Once all of the sequences were classified, a kappa statistic was calculated to determine how reliably the HMMs classified lifting sequences that were generated from a particular model.

B. Second Simulation

The second simulation study assessed whether the HMMs can reliably identify lifting sequences to the appropriate model when some of the lifting sequences are intentionally misclassified. A small number of simulated lifting sequences were switched between the groups to create intentionally misclassified sequences. For example, lifting sequences generated from the CLBP HMM was labeled as lifting sequence generated from control HMM and vice versa. A jackknife method was used to train and test the HMMs. This method excluded a single sequence and retrained the control and CLBP models with the remaining sequences. Initial estimates of the transition and token probabilities were recalculated for each retraining. After training, the excluded lifting sequence was classified to one of the HMMs based on the likelihood probabilities. This process continued until all of the lifting sequences were tested.

The sample size of the simulated sequences in the second simulation was chosen as 108 (54 CLBP simulated sequences and 54 control simulated sequences) to approximately match the data of the clinical study. The number of lifts in the sequences and the starting state of the simulated sequences was determined using a random number generator in MatLab. Fifty percent of the simulated sequences generated from the CLBP HMM were randomly chosen to have a sequence length that varied from 6 lifts to 20 lifts and the remaining 50% were randomly chosen to have a length that varied from 21 lifts to 80 lifts. All of the simulated sequences generated from the control HMM were randomly assigned to a length that varied from 21 lifts to 80 lifts. The number of intentionally misclassified sequences was varied equally between the groups (e.g. 1 simulated sequence generated from the control HMM and 1 simulated sequence from the CLBP HMM were both misclassified) from 4 to 64 (4% to 50% of the total sample size) in increments of 5 (9%). At each increment in the number of intentionally misclassified sequences, all of the sequences were individually tested against both HMMs using the jackknife method and classified to a HMM based on the likelihood probability. A kappa statistic assessed reliability.

III. Results

In both simulation studies, kappa was > 0.8 indicating that the two HMMs are highly reliable and can be applied to the clinical data. The results are described below.

A. First simulation

In the first simulation study, the number of lifts and sample size of the simulated sequences were varied in all three trials. The starting state varied between trials but not within the three trials. Kappa was calculated for each group of simulated sequences that had equal length and started in the same state. For instance, simulated sequences generated from the CLBP HMM that started in state 1 and contained 6 lifts were compared to simulated sequences generated from the control HMM that started in state 1 and contained 6 lifts. In all three trials, kappa was greater than 0.8 for all sample sizes and sequence lengths. Figure 2 shows a plot of kappa versus number of lifts in the sequence for each sample size when the sequences started in state 3. These results indicate that the HMMs are highly reliable for different sample sizes when equal length sequences that all start in the same state are compared.

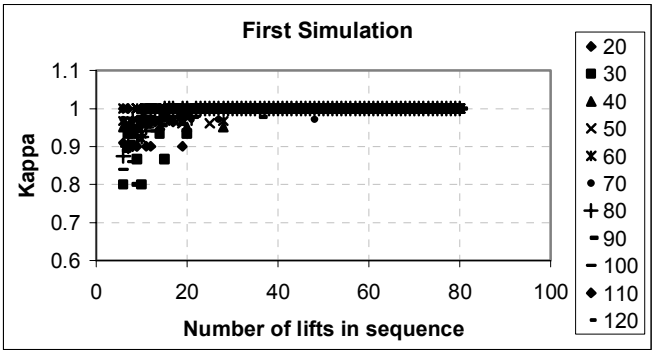

Figure 2: Kappa plotted versus sequence length for each sample size when the simulated sequences started in state 3. Symbols indicate sample size.

Since the lifting sequences did not all start in the same state nor were the length of the lifting sequences equal in the clinical data, the first simulation was repeated on data that approximately match the clinical data. The simulated

sequences generated in this version of the first simulation study were tested and classified to a model based on a likelihood probability. Kappa assessed reliability.

The results of the second analysis of the first simulation study found one classification error resulting in a kappa of 0.98. The high reliability indicates that the HMMs are highly reliable when the simulated sequences approximately match the clinical data.

B. Second Simulation

In the second simulation, a percentage of the data was misclassified and the HMMs were trained with the misclassified data. The jackknife method was used to test and retrain the HMMs. Kappa determined reliability at each incremental increase in the percentage of misclassified sequences. The HMMs were found to have excellent reliability when 4% to 41% of the data was intentionally misclassified as shown in Figure 3. This statistic corresponded to classification error of 2 to 4 simulated sequences. At 50% of the data intentionally misclassified, the HMMs showed fair reliability (kappa = 0.26) and the number of classification errors was 40 sequences.

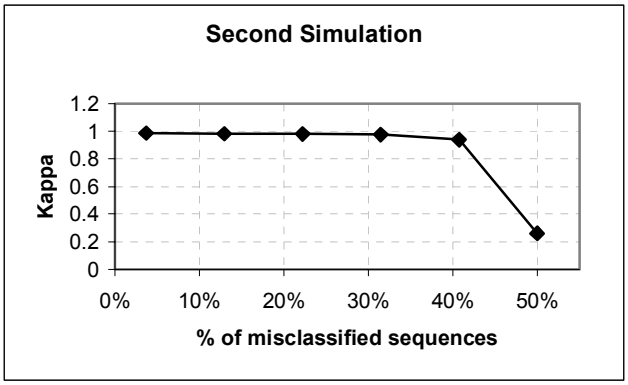

Figure 3: Kappa plotted versus % of misclassified sequences

IV. Discussion

In both of the simulation studies, kappa indicated that the HMMs have excellent reliability (kappa > 0.8). The high reliability implies that the two HMMs are describing different lifting patterns during the repetitive lifting task. This result is consistent with previous research that has shown temporal difference in lifting patterns between control and CLBP subjects [3]-[5].

The jackknife method used to train the HMMs in the second simulation study permitted classification of each simulated lifting sequence to one of two models without introducing bias associated with classifying sequences that were used to train the HMM. The method avoided bias in the training and testing of the HMMs by excluding a single test sequence and retraining the HMMs with the remaining data. The test sequence was tested against the retrained HMMs and classified to one of the models. The HMMs trained with the jackknife approach were found reliable, indicating that the jackknife method is a valid technique to train HMMs.

The HMMs were found to have excellent reliability in both simulation studies suggesting that these models may reliable identify CLBP subjects whose time series of lifting patterns are more similar to control subjects than other CLBP subjects. In this study, the HMMs of both groups were ergodic topologies. Future work will focus on developing HMMs with simpler topologies that incorporate the dynamic structure of the data without reducing reliability.

V. Conclusion

This study evaluated the reliability of HMMs that describe the lifting patterns of control and CLBP subjects during a repetitive lifting task using two simulation studies. The first simulation determined that the control and CLBP HMM can reliably identify simulated sequences that were generated from the HMMs. The second simulation determined that the HMMs can reliably identify simulated sequences to the appropriate model when the models are trained with intentionally misclassified data. The jackknife method was used as an alternative approach to training and testing the HMMs in the second simulation and was found to be a reliable technique. Since a kappa statistic indicated that the HMMs were highly reliable in both simulation studies, the models can be used to identify sub-groups of subjects, based on lifting patterns used during a repetitive lifting task.

References

[1] Turk D.C. and Rudy T.E. Towards a comprehensive assessment of chronic pain patients. Behavioral Research Therapy. 1987; 25 (4) 237-249.

[2] Turk D.C. and Rudy T. E. The robustness of an empirically derived taxonomy of chronic pain patients. Pain 1990; 42:27-35.

[3] Rudy T.E., Boston J.R., Lieber S.J., Kubinski J.A., and Stacey B.R. Body motion during repetitive isodynamic lifting: a comparative study of normal subjects and low-back pain patients. Pain, 2003; 105:319-326.

[4] Slaboda J.C., Boston J.R., Rudy T.E., Lieber S.J. and Rasetshwane D.M. The use of splines to calculate jerk for a repetitive lifting task involving chronic lower back patients. IEEE Trans Neur Sys and Rehabil Eng. 2005; 13 (3): 406-414.

[5] Vasko R.C., El-Jaroudi A. and Boston J.R. Application of hidden Markov model topology estimation to repetitive lifting data. Acoustics, Speech, and Signal Processing, ICASSP-97., IEEE International Conference. 1997; 5: 4073-4076.

[6] Eye A., and Schuster C. Regression analysis for social sciences. San Diego, CA: Academic Press, 1998.

[7] Milligan G.W. and Cooper M.C. An examination of procedures for determining the number of clusters in a data set. Psychometrika. 1980; 50 (2): 159-179.

[8] Rabiner L.R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 1973; 77 (2):257-286.

[9] McGinn T., Wyer P.C., Newman T.B., Keitz S., Leipzig R. and Guyatt G. Tips for learners of evidence-based medicine: 3. Measures of observer variability (kappa statistic). Canadian Med Assoc J 2004; 171(11): 1369-1373.