Latent Structure Models and their Applications in Informatics

John C. Paolillo, Instructor

Tuesdays, 6:00-8:30 PM, WH 119

Office	Hours
LI 030A	T 4:30-6:00, R 4:00-5:30
Eigenman 912	TBA

Quick Links

- PDF version of syllabus Δ (current 8-30-2007)
- Course wiki

Course Description

Many problems in informatics involve the analysis of large numbers of variables in similarly large numbers of contexts: the data may be luminescence values in a series of assays from a microarray experiment, receipts in a database of purchases, term counts in a set of documents, base sequences in a scomparison set of genomes, links in a set of web pages, or many other things. Among the many data mining techniques available, a family of statistical models known as Latent Structure Models (LSMs) hold a particular distinction, chief representatives of which are Principal Components Analysis and Factor Analysis. These are well-understood statistical models for analysis of high-dimensional data, there are many examples of their application to a wide range of problems, and they can scale to very large data sets. Recent literature in statistics also extends these models to a wide range of circumstances.

This course is a seminar-style course, focusing on the application of Latent Structure Models to research problems in Informatics. Readings for the course are drawn from the current literature on these models and their applications. Depending on students background and interests, readings will be covered that pertain to any of the following topics: mathematical underpinnings of LSMs; computational implementation of LSMs; analysis and interpretation of LSMs; applications of LSMs to specific problems in Bioinformatics, Chemical Informatics, HCI/Usability, Social Informatics, Information Retrieval, Natural Language Processing, etc. Students are expected to participate in and lead discussion of research articles discussing these models and their application to research in areas of their own interest. In addition, students will undertake a research project in which a LSM is applied to a research question, to be determined by the student in consultation with the instructor.

Prerequisites

This course has no formal prerequisites, though one or more of the following are likely to be useful:

knowledge of elementary statistics, matrix/linear algebra, background in an informatics subject area (bioinformatics, HCI, Information Retrieval, data mining, etc.)

Course Objectives

This course aims to develop a critical appreciation of Latent Structure Models and their application to Informatics. At the end of this course, students should be able to

- recognize research problems that may and may not benefit from the application of LSMs
- recognize the main types of LSMs
- identify how certain types of LSMs might be examined empirically for a given application
- conduct and interpret a LSM analysis of an original data set selected to address an meaningful research problem

Textbook(s)

The following titles provide the central set of readings for this course. Basilevsky (1994) is required; parts of Bartholomew and Knott (1999) may be used as well. There are similarities but they are non-overlapping.

- Basilevsky, A. 1994. Statistical Factor Analysis and Related Methods: Theory and Application. New York: Wiley.
- Bartholomew, D.J.; and M. Knott. 1999. Latent Variable Models and Factor Analysis. London: Arnold.

A number of readings are indicated in the course schedule; normally these will be available online, either on the WWW or through the IU Library system. In addition, students are responsible for identifying in advance and making available 4 relevant readings for discussion by the class. Note: for the print books, it is wise to comparison-shop in order to get the best availability and lowest prices.

Requirements

The course aims are achieved through:

- Readings and discussion of current research about or employing LSMs
- Directed research projects applying LSMs

The graded work in this course comprises

- five article presentations worth five points each
- two research progress reports worth five points each
- a final presentation worth 15 points
- a final paper or project worth 50 points

The research progress reports, the final presentation and the final paper are intended to provide support for developing an in-depth research project that spans the entire semester; the topic of this

research will be chosen by each student in consultation with the instructor.

Calendar

The following calendar outlines our activities for the semester. This portion of the syllabus, in particular, may change, so please check it regularly. You are responsible for readings on the day indicated.

Date	Topics and Readings	Assignments
Aug 28		
Sept 4		
Sept 11		
Sept 18		
Sept 25		
Oct 2		
Oct 9		
Oct 16		
Oct 23		
Oct 30		
Nov 6		
Nov 13		
Nov 20		
Nov 27	Presentations	Presentations
Dec 4	Presentations	Final paper due

Readings

These are a few of the suggested readings

- Bartholomew, D.J. 2004. Latent variable modeling: Old and new approaches. http://statwww.epfl.ch/teaching/3eCycleRomand/bartholomew.talk.pdf
- Bartholomew, D.J. 2004. Three faces of factor analysis. http://www.fa100.info/bartholomew.pdf
- Bartholomew, D.J.; and M. Knott. 1999. Latent Variable Models and Factor Analysis. London: Arnold.
- Basilevsky, A. 1994. Statistical Factor Analysis and Related Methods: Theory and Application. New York: Wiley.
- Berry, M.; S.T. Dumais; and G.W. O'Brien. 1995. Using matrix algebra for intelligent information retrieval. SIAM Review, 37:573-595.
- Berry, M.W.; and G.W. Stuart. 2003. A comprehensive whole-grnome bacterial phylogeny using correlated peptide motifs defined in a high-dimensional vector space. Journal of Bioinformatics and Computational Biology, 1.3: 475-493.
- Biber, D. 1988. Variation across Speech and Writing. Cambridge: Cambridge University Press.
- Biber, D. 1995. Dimensions of Register Variation: A Cross-Linguistic Comparison. Cambridge:

Cambridge University Press.
- Borg, I.; and P. Groenen. 1997. Matrix Algebra for MDS. Chapter 7 of Modern Multidimensional Scaling: Theory and Applications, 109-134. Berlin: Springer.
- Burgess, C.; K. Livesay; and K. Lund. 1998. Explorations in context space: words, sen-tences, discourse. Discourse Processes, 25: 211-257.
- Burrows, J.F. 1992. Computers and the study of literature. In C. Butler, ed., Computers and Written Texts, 167-204. Oxford: Blackwell.
- Dumais, S.T. 2003. Data-driven approaches to information access. Cognitive Science, 27:491-524.
- Foltz, P. W.; W. Kintsch; and T.K. Landauer. 1998. The measurement of textual coherence with Latent Semantic Analysis. Discourse Processes, 25:285-307.
- Hoyle, R.H.; and J.L. Duvall. 2004. Determining the number of factors in exploratory and confirmatory factor analysis. In D. Kaplan, ed., The Sage Handbook of Quantitative Methodology for the Social Sciences, 301-315. Thousand Oaks: Sage.
- Jordan, M.I. 2004. Graphical models. Statistical Science, 19.1: 40-155.
- Lan, H.; J.P. Stoehr; S.T.Nadler; K.L. Schueler; B.S. Yandell; and A.D. Attie. 2003. Dimension reduction for mapping mRNA abundance as quantitative traits. Genetics, 164: 1607-1614.
- Kim, Y.-S.; J.-H. Chang, and B.-T. Zhang. 2003. An empirical study on dimensionality optimization in text mining for linguistic knowledge acquisition. Lecture Notes in Artificial Intelligence, 2637, 111–116.
- Kintsch, W.; and A. Bowles. 2002. Metaphor comprehension: What makes a metaphor difficult to understand? Metaphor and Symbol, 17:249-262.
- Kintsch, W. 2001. Predication. Cognitive Science 25:173-202.
- Landauer, T. K.; and S.T. Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.
- Landauer, T. K.; P.W. Foltz; and D. Laham. 1998. Introduction to Latent Semantic Analysis. Discourse Processes, 25:259-284.
- Lavie, T.; and N. Tractinsky. 2004. Assessing dimensions of perceived visual aesthetics of websites. International Journal of Human-Computer Studies, 60: 269-298.
- Lan, H.; J.P. Stoehr; S.T.Nadler; K.L. Schueler; B.S. Yandell; and A.D. Attie. 2003. Dimension reduction for mapping mRNA abundance as quantitative traits. Genetics, 164: 1607-1614.
- Marcoulides, G.A.; and I. Moustaki. 2002. Latent Variable and Latent Structure Models. Mahwah: Lawrence Erlbaum.
- Miller, L.A.; K.M. Stanney; and W. Wooten. 1997. Development and evaluation of the Windows Computer Experience Questionnaire (WCEQ). International Journal of Human-Computer Interaction, 9.3:201-212.
- Morrison, D.A.; and J.T. Ellis. 2003. The design and analysis of microarray experiments: applications in parasitology. DNA and Cell Biology, 22.6: 357-394.
- Paolillo, J.C. 2004. Latent structure analysis: Semantic or syntactic? manuscript. http:ella/slia.indiana.edu/~paolillo/research/paolilloICON2004.pdf
- Rehder, B.; M.E. Schreiner; M.B. Wolfe; D. Laham; T.K. Landauer; and W. Kintsch. 1998. Using Latent Semantic Analysis to assess knowledge: some technical con-siderations. Discourse Processes, 25, 337-354.
- van Schaik, P.; and J. Ling. 2003. Using on-line surveys to measure three key constructs of the quality of human-computer interaction in web sites: psychometric properties and implications. International Journal of Human-Computer Studies, 59: 545-567
- Sigley, R. 1997. Text categories and where you can stick them: a crude formality index. International Journal of Corpus Linguistics, 2:199-238.
- Simon, A.F.; and M. Xenos. 2004. Dimensional reduction of word-frequency data as a

substitute for intersubjective content analysis. Political Analysis, 12:63–75.
- Tacq, J. 1997. FactorAnalysis: The investigation of marital adjustement. Chapter 9 of Multivariate Analysis Techniques in Social Science Research. Thousand Oaks: Sage.
- Zhang, N.L. 2004. Hierarchical Latent Class Models for Cluster Analysis. Journal of Machine Learning Research 5(Jun):697--723, 2004. http://www.jmlr.org/papers/volume5/zhang04a/zhang04a.pdf

Additional Resources

- http://www.fa100.info/spkr.htm
- http://ftp.isds.duke.edu/WorkingPapers/98-38.pdf
- http://www.ai.mit.edu/projects/jmlr/

(more to be added later)

Assignments

- Lead Discussion. (5) Throughout the semester. Each student will be responsible for leading discussion connected with one of the selected articles, on five days throughout the course, selected from recommended readings or other material relevant to the student's interests (to be cleared by the instructor). The selected readings must be publicized well in advance of the discussion date (and made available to the class if necessary), so that we all have an opportunity to read the material prior to discussing it.
- Research report. (2) September 27 and November 1. There are two research reports, intended to provide you an opportunity to discuss the current state of your research while it is still in progress so that you may solicit input about the challenges you face, and ways to resolve them. The first will report on the initial conception of your research, the second will report on the state of the research in progress. All issues relevant to the conduct of the research are permissible topics of discussion. As we will need to handle several research reports in any one class, these will also need to be reasonably short.
- Final presentation. (1) December 6. You will present on the state of your research project during the final week of class.
- Final Paper/Project. (1) Due December 10. A final written version of your research for the semester is due after the last class. As I am likely to attend ICON 2004 in Hyderabad, India, I will need to have your final papers promptly by the end of the last week of classes, so that I can submit your grades without delay.

Assignments in this class are evaluated according to the following table.

Assignment type	value	number	Total
Lead Discussion	5%	5	25%
Research Report	5%	2	10%
Presentation	15%	1	15%
Final Paper	50%	1	50%

Course Requirements

To receive a passing grade in this course, you must turn in all of the assignments and the term project and do your presentations. You cannot pass this course without doing all of the assigned work, however, turning in all of the work is not a guarantee that you will pass the course. Grades of I (Incomplete) may be assigned in this course after discussion with the instructor, but, depending on the circumstances, there will be a penalty applied at the discretion of the instructor.

All papers and assignments must be submitted on the dates specified in this syllabus. If you cannot submit an assignment or cannot deliver a presentation on the date it is due, it is your responsibility to discuss your situation with the instructor, preferably in advance. Given that your reasons or problems are legitimate, arrangements for the completion of the outstanding work can be made; this will occur, however, at the discretion of the instructor. There will be a penalty for work turned in after the assigned date, and this will also be applied at the discretion of the instructors.

Your written, web-based, and oral work will be evaluated according to four criteria; it must:

- Be clearly written, marked up, and/or presented, and checked for spelling and grammar;
- Demonstrate a degree of insight into the concepts, issues, and trends in both the areas you investigate in the assignments and in the course content;
- Demonstrate a degree of originality in your reviews, analyses and projects; and
- Display some familiarity with the appropriate current and/or classic literatures where appropriate.

Borderline grades will be decided (up or down) on the basis of class contributions and participation throughout the semester. Indiana University and School of Informatics policies on academic dishonesty will be followed. Students found to be engaging in plagiarism, cheating, and other types of dishonesty will receive an F for the course. As a rule of thumb, when in doubt, cite the source!

The following definitions of letter grades have been suggested to Informatics Faculty as an aid in evaluation of academic performance and to assist students by giving them an understanding of the grading standards of the School of Library and Information Science:

Grade	GPA	Meaning
A	4.0	Outstanding achievement. Student performance demonstrates full command of he course materials and evinces a high level of originality and/or creativity that far surpasses course expectations
A-	3.7	Excellent achievement. Student performance demonstrates thorough knowledge of the course materials and exceeds course expectations by completing all requirements in a superior manner
B+	3.3	Very good work. Student performance demonstrates above-average comprehension of the course materials and exceeds course expectations on all tasks as defined in the course syllabus
B	3.0	Good work. Student performance meets designated course expectations, demonstrates understanding of the course materials and is at an acceptable level
B-	2.7	Marginal work. Student performance demonstrates incomplete understanding of course materials.
C+	2.3	Unsatisfactory work. Student performance demonstrates incomplete and inadequate understanding of course materials

C	2.0	
C-	1.7	Unacceptable work. Course work performed at this level will not count toward the MLS or MIS degree. For the course to count towards the degree, the student must repeat the course with a passing grade.
D+	1.3	
D	1.0	
D-	0.7	
F	0.0	Failing. Student may continue in program only with permission of the Dean.

This page maintained by John C. Paolillo