Sparse Factor Analysis for Cognitive Tutoring

Andrew E. Waters, Andrew Lan, Christoph Studer, and Richard G. Baraniuk
Rice University
e-mail: {andrew.e.waters, sl29, studer, richb}@rice.edu

We currently live in an era of vast disparity in regards to the educational opportunities available to people worldwide. This disparity is felt especially among underprivileged classes and those living in developing countries. One of the great challenges of the 21st century is to create the infrastructure necessary to overcome this disparity. One very attractive avenue for ameliorating this condition is the development of cognitive education tutors based on machine-learning concepts [2]. These systems would allow for high-quality education materials to be distributed to a very broad audience that does not have access to universities or human instructors. While a significant amount of work has been devoted during the past decade toward developing cognitive tutors, current systems are still quite limited and fall into one of two broad categories:

- *Rule-based/expert-designed systems*: These systems are generally of very high quality. However, they also require enormous initial investment to create and remain fixed after development.
- *Automated systems based on statistical models*: These systems offer the promise of very affordable systems that can adapt to various conditions over time. However, current systems generally use unsophisticated machine learning techniques, which limit their efficacy.

We envision a statistically-minded cognitive tutor that is able to learn about the student as the student learns about the subject material being taught. This approach would allow the system to naturally assess which knowledge areas the student understands well, as well as which areas are still problematic. When deficiencies are identified, the system can propose corrective action, e.g., suggest additional reading material. Moreover, when the student demonstrates sufficient mastery of the material, the system can proceed to teach new concepts.

Realizing such a cognitive tutoring system presents a number of fascinating challenges. For example, how should we model the target knowledge in a statistically principled manner? How can we reliably estimate a student's understanding of the various subject material? Furthermore, given a database of potential practice problems, how can we determine which problems are relevant for teaching and evaluating certain knowledge? Finally, can our system make reasonable predictions of student ability, such as determining whether or not a student will complete a given problem successfully?

To address these challenges, we propose a novel statistical framework based on Bayesian latent factor analysis [3]. To this end, we assume that the knowledge base is decomposable into a set of latent knowledge concepts that will be learned by the student. As an example, an introductory calculus course might have latent concepts such as integration by parts, differentiation of sinusoidal functions, l'Hôpital's rule, etc. In particular, given q latent concepts, n students, and p questions, we model the probability of student i answering question j correctly via the following probit regression model:

$$p(y_{i,j}) = \Phi\big(\mu_j + \mathbf{\Lambda}_j^T \mathbf{F}_i + e_{i,j}\big)\,.$$

Here, μ_j is a mean shift for question j (corresponding to the intrinsic difficulty of the jth question), $\mathbf{\Lambda}_j \in \mathbb{R}^q$ is the factor loading for question j, $\mathbf{F}_i \in \mathbb{R}^q$ is the concept mastery vector for student i, and $e_{i,j}$ models other random statistical effects. The function $\Phi(\cdot)$ is the probit link function.

Given a binary-valued observation matrix $\mathbf{Y} \in \{0,1\}^{p\times n}$ indicating whether the answers of the n students to all p questions are correct or not, we wish to infer the factor loadings $\mathbf{\Lambda}_j$ and the student concept mastery $\mathbf{F}_i$ for all questions and all students. Inference is accomplished in a Bayesian fashion by proposing analytically tractable likelihood functions and prior probabilities on the random variables of interest. We further assume sparsifying priors on the factor loadings $\mathbf{\Lambda}_j$, consistent with the philosophy that most questions should only relate to a small number of the available latent concepts. By using techniques from Bayesian data augmentation [1] we develop a Markov chain Monte-Carlo (MCMC) method based on Gibbs' sampling to compute posterior distributions for the latent variables of interest. This approach enables us to both determine a model for how concepts intersect with questions, but also allow us to determine student understanding of the material.

We demonstrate the performance and efficacy of our approach on both synthetic data, as well as on real-world data collected from actual students enrolled in a signal-processing course at Rice University. Our results demonstrate that our method is useful in determining both the latent feature structure of the data, as well as the student understanding of the knowledge concepts. This framework paves the way for designing novel algorithms that can improve the state-of-the-art in cognitive tutoring systems.

References

[1] J.H. Albert and S. Chib. Bayesian analysis of binary and polychotomous data. *J. of the American Stat. Soc.*, 88(422):669–679, 1993.

[2] J.A. Kulik. Meta-analytic studies of findings on computer-based instruction. *Technology assessment in education and training*, pages 9–33, 1994.

[3] I. Pournara and L. Wernisch. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, 8(1):61, 2007.