Examining Patterns of Sexual Behaviors over Two Years among Runaway Youths

Juwon Song
Unversity of California, Los Angeles, Department of Biostatistics
10920 Wilshire Blvd. Suite 350
Los Angeles, CA, United States
jwsong@ucla.edu

Thomas R. Belin
Unversity of California, Los Angeles, Department of Biostatistics
CHS 51-267
Los Angeles, CA, United States
tbelin@MEDNET.ucla.edu

Mary Jane Rotheram-Borus
Unversity of California, Los Angeles, Department of Psychiatry
10920 Wilshire Blvd. Suite 350
Los Angeles, CA, United States
rotheram@ucla.edu

In a study of runaway youths, 311 participants were recruited from four shelters in New York and followed over two years. One hundred sixty seven participants from two shelters received an intervention to reduce HIV risk acts and 144 participants from two shelters served as a control group. Since randomization occurred in shelters, some baseline characteristics between the intervention and the control groups were unbalanced, and this difficulty was handled by using subclassification on the propensity score (Song, et. al., 2001).

One purpose of the study is to investigate their patterns of sexual behaviors over time. For example, some participants' number of risky sexual acts was decreased over time, some showed increased numbers over time, and others showed random fluctuations. This indicates that there are several groups of participants showing different trajectories over time. The purpose of this presentation is to cluster participants with similar patterns of sexual behaviors.

A model-based clustering technique is adopted for this purpose. It addresses cluster analysis as an analysis of mixtures of multivariate distribution (Wolfe 1970, Fraley and Raftery 2000). Let's denote that $f_1(Y,q_1)$, $f_2(Y,q_2)$, .., $f_g(Y,q_g)$ are g probability distributions defined on a p-dimensional space of random vectors Y and parameters q_1, q_2, .., q_g, and p_1, p_2, .., p_g are probabilities that observations belong

to corresponding probability distributions. Then, the likelihood of the mixture model with g components is

$$L(\theta_1,\theta_2,...,\theta_g;\pi_1,\pi_2,...,\pi_g|Y)=\prod_{i=1}^{n}\sum_{k=1}^{g}\pi_k f(y_i|\theta_k) \text{ for independent } y_i,\ i=1,...,n.$$

Fraley and Raftery (2000) describe the EM algorithm (Dempster, Laird and Rubin 1977) to obtain MLE of θ and π. This is a special case of MLE for mixed continuous and categorical data, considering π as missing categorical variables (Little and Schluchter 1985). Moreover, it can be extended to the case with partially observed Y (Schafer 1997, Chapter 9.4.1).

We apply this technique to the study of runaway youths and examine whether we can successfully cluster different patterns of sexual behaviors over two years. Since there are often missing follow-ups, Y is partially observed in this data set, and we use the EM algorithm with partially observed Y and completely missing π. In this approach, choosing an appropriate number of cluster, g, corresponds to choose an appropriate statistical model. Therefore, we use Bayesian model selection (Fraley and Raftery 2000) to choose an appropriate number of clusters.

REFERENCE

Dempster, A.P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood Estimation from Incomplete Data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, **39**, 1-38.

Fraley, C. and Raftery A. E. (2000). Model-Based Clustering, Discriminant Analysis, and Density Estimation, Technical Report no. 380, Department of Statistics, University of Washington.

Little, R. J. A. and Schluchter, M.D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values, *Biometrika*, **72**, 497-512.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data, Chapman and Hall, London.

Song, J., Belin, T. B., Lee, M. B., Gao, X., and Rotheram-Borus, M. J. (2001). Handling Baseline Differences and Missing Items in a Longitudinal Study about Reductions in HIV Risk among Runaway Youths, submitted manuscript.

Wolfe, J. H. (1970) Pattern Clustering by Multivariate Mixture Analysis. *Multivariate Behavioral Research*. **5**, 329-350.

RESUME

Dans le cadre d'une étude de jeunes fugueurs, 311 participants ont été recrutés dans quatre refuges. Cette étude a pour objectif d'analyser les types de comportement sexuel qu'ils adoptent dans le temps. Nous appliquons une technique de groupement sur modèle mettant en œuvre l'algorithme EM pour déterminer s'il est possible de grouper avec succès différents types de comportement sexuel sur une période de deux ans.