Stephen E. Fienberg's Contributions to Categorical Data Analysis and the Social Sciences

Edoardo M. Airoldi

This year at JSM, I had the privilege of providing an outlook on Steve's classic work on categorical data analysis and applications to the social sciences and to highlight some recent developments it inspired. This column recapitulates a few points of the conversation.

Since the beginning of his career at Harvard, Steve has been pioneering work on the geometry of contingency tables. He first formalized the problem by conceptualizing 2 × 2 tables as points in a reference tetrahedron—shown in Figure 1, left panel. In a 1970 paper, he used this formalism to characterize parametric spaces that arise in statistical problems, such as the surfaces of constant association, and to conclude that the (spiral) path to convergence of the iterative proportional fitting procedure always lies on the surface of constant association determined by the table used to initialize the procedure—illustrated in Figure 1, right panel. Steve also extended this formalism, and a number of key results, to r × c contingency tables in a follow-up paper, which appeared earlier. (Ask Steve for a good story.)

Following this thread, we recently developed an interactive Java applet, available for download at *www.fas.harvard.edu/~airoldi,* to help students and instructors explore the geometry of 2 × 2 contingency tables. For example, the screen shot in Figure 2 illustrates Simpson's paradox: two tables that correspond to sub-group analyses have odds greater than one and lie on the same side of the surface of independence (displayed in green), while the table that corresponds to the overall analysis (the dot in the middle) has odds less than one and lies on the opposite side of the surface of independence.

Steve's interest in the algebraic and polyhedral geometry of problems involving contingency tables—ill-posed inverse problems and log-linear models—have inspired a steady stream of graduate research and resulted in substantial contributions to many application areas, from strategies for analyzing the U.S. Census, the National Long Term Care Survey, and other national surveys, to issues of privacy and disclosure limitation in medical and genomic databases, to social network analysis.

During the session, Judith Tanur discussed Steve's work in the social sciences, broadly. Part of my talk summarized Steve's work on mixed membership models, specifically with application to disability survey data and the analysis of text. I reviewed Steve's clear characterization of the four levels of assumptions underlying this class of models. These models have gained tremendous traction in recent years, and a handbook of mixed membership models, curated by Steve, is scheduled for publication in 2014 by CRC Press.

I concluded with a selection of Steve's methodological contributions to social network analysis.

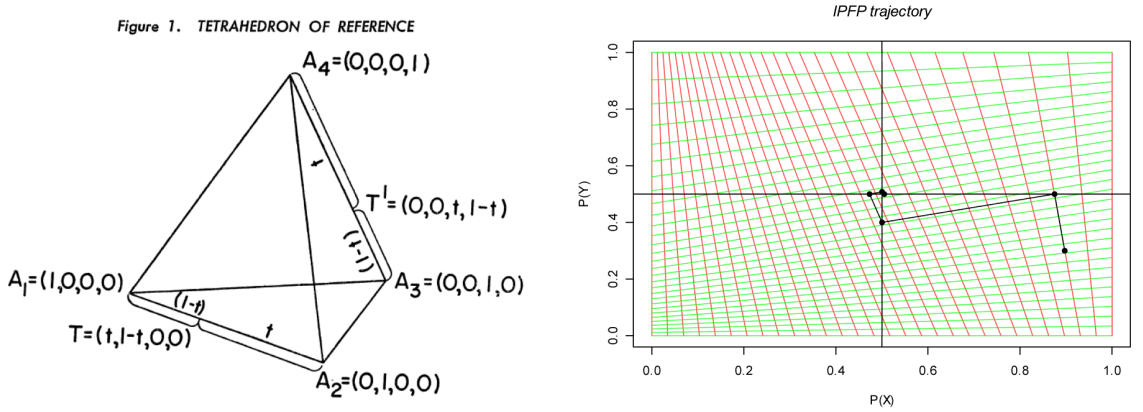

Figure 1: The original illustration of the reference tetrahedron (left panel) and an illustration of the path to convergence instantiated by iterative proportional fitting procedure (right panel)

Among my favorite papers, I mentioned a timeless analysis of Sampson's monastery data, which provides the basis for dynamic theories of social failure in isolated communities and remains the most sophisticated analysis of those data to date. Also, an analysis of Stanley Milgram's experiments, which considers alternative models for incomplete letter chains and ultimately casts doubts on the popular claim of "six degrees of separation." I also noted how Steve's analysis is quite relevant today, as modern studies that have revisited Milgram's experiments on social media platforms most often fell victim to the same shortcomings he pointed out in 1975. I reviewed results in a recent paper I co-authored with Steve that combines mixed membership models and stochastic block-models and develops a "nested" variational inference strategy for analyzing large, sparse networks. I concluded by highlighting a recent paper that uses geometric arguments to provide insight into the conditions that guarantee the existence of a maximum likelihood estimator for a popular model of network data—the β-model.

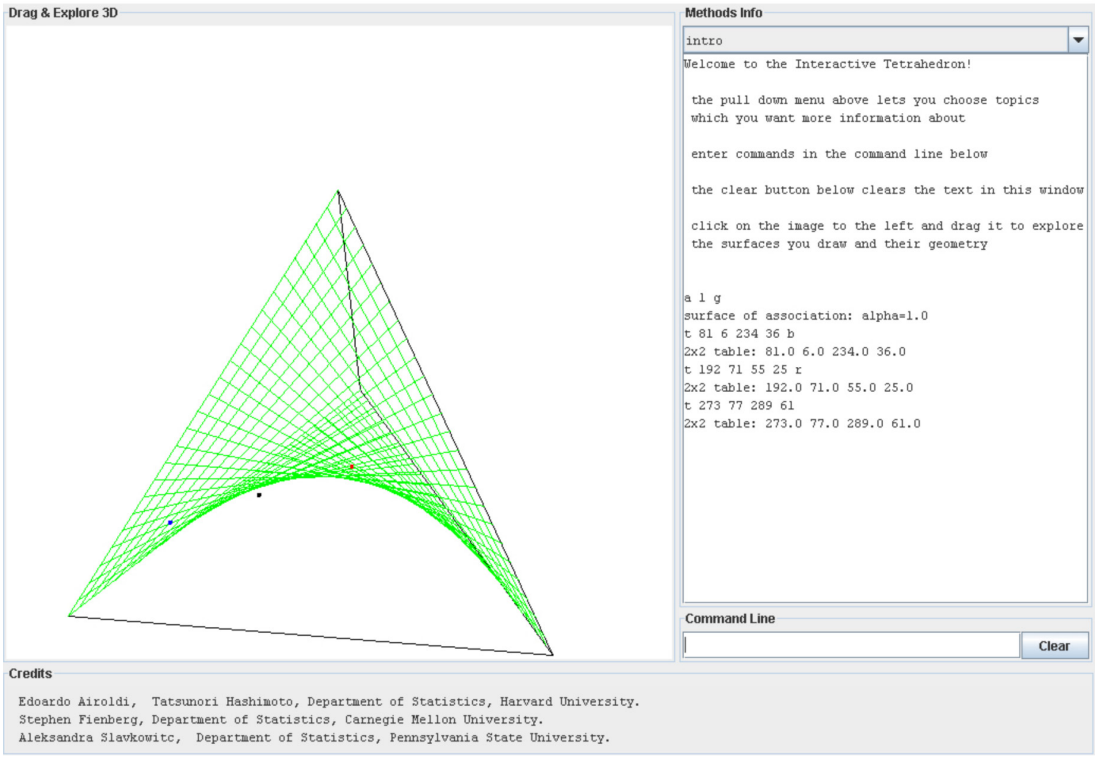

Figure 2: An illustration of Simpson's paradox using the 3D tetrahedron applet. Famously, for example, it is possible for a graduate school to have a favorable overall admission rate for females, while having unfavorable admission rates for females in each department at the same time.

The session in honor of Steve's 70th birthday included contributions by Ed George, Judith Tanur, and Sesa Slavkovic, as well as discussions by Jim Berger and Steven Stigler—both informative and highly entertaining. Throughout the session, there was laughter, as well as some tears and a hug. Eventually, we made our way through Montréal to red wine and a good meal. Thank you, Steve!

Further Reading

Airoldi, E. M. 2006. Bayesian mixed membership models of complex and evolving networks. PhD thesis, School of Computer Science, Carnegie Mellon University

Airoldi, E. M., D. M. Blei, S. E. Fienberg, and E. P. Xing. 2008. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014.

Airoldi, E. M., E. A. Erosheva, S. E. Fienberg, C. Joutard, T. M. Love, and S. Srhingarpure. 2010. Reconceptualizing the classification of PNAS articles. *Proceedings of the National Academy of Sciences*.

Bickel, P. J., E. A. Hammel, and J. W. O'Connell. 1975. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175):398–404.

Bishop, Y., S. E. Fienberg, and P. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press.

Bishop, Y. M., S. E. Fienberg, and P. W. Holland. 2007. *Discrete Multivariate Analysis: Theory and Practice*. Springer, second edition,.

Blei, D. M., A. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Deming, W. E. and F. F. Stephan. 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11(4):427–444.

Dobra, A. 2001. Statistical tools for disclosure limitation in multi-way contingency tables. PhD thesis, Carnegie Mellon University, Department of Statistics.

About the Author

Edoardo M. Airoldi is an associate professor of statistics at Harvard University and an associate faculty member at the Broad Institute of MIT & Harvard. He earned his PhD in computer science from Carnegie Mellon University and worked with Stephen Fienberg and Kathleen Carley at Princeton University. He is the recipient of an NSF Career Award, an Alfred P. Sloan Research Fellowship, and several outstanding paper awards, including the John Van Ryzin and Thomas R. Ten Have awards.

Erosheva, E. A. 2002. Grade of membership and latent structure models with application to disability survey data. PhD thesis, Carnegie Mellon University, Department of Statistics.

Erosheva, E. A., S. E. Fienberg, and J. Lafferty. 2004. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 97(22):11885–11892.

Erosheva, E. A., S. E. Fienberg, and C. Joutard. 2007. Describing disability through individual-level mixture models for multivariate binary data. *Annals of Applied Statistics*, 1:502–537.

Fienberg, S. E. 1968. The geometry of an r × c contingency table. *Annals of Mathematical Statistics*, 39:1186–90.

Fienberg, S. E. and J. P. Gilbert. 1970. The geometry of a two by two contingency table. *Journal of the American Statistical Association*, 65:694–701.

Fienberg, S. E. and S. K. Lee. 1975. On small world statistics. *Psychometrika*, 40(2):219–228.

Fienberg, S. E., M. M. Meyer, and S. Wasserman 1985. Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80:51–67.

Milgram, S. 1967. The small world phenomenon. *Psychology Today*, 1(61).

Nowicki, K. and T. A. B. Snijders. 2001. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96:1077–1087.

Pritchard, J., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959.

Rinaldo, A. 2005. On Maximum Likelihood Estimation for Contingency Tables. PhD thesis, Carnegie Mellon University, Department of Statistics.

Rinaldo, A., S. Petrovic, and S. E. Fienberg. 2013. Maximum likelihood estimation in the β-model. *Annals of Statistics*, 41(3).

Sampson, F. S. 1968. A Novitiate in a period of change: An experimental and case study of social relationships. PhD thesis, Cornell University.

Slavkovic, A. B. 2004. Statistical Disclosure Limitation Beyond the Margins. PhD thesis, Department of Statistics, Carnegie Mellon University.

Straf, M. L. and J. M. Tanur. 2013. A conversation with Stephen E. Fienberg. *Statistical Science*, 28(3):447–463.

Woodbury, M. A., J. Clive, and A. Garson. 1978. Mathematical typology: Grade of membership technique for obtaining disease definition. *Computational Biomedical Research*, 11(3):277–298.