Final Report
2004-05 Program on
Latent Variable Models in the Social Sciences

June 13, 2006

1. Introduction, Motivation and Initial Ideas

The purpose of the 2004–05 SAMSI program on the Latent Variable Models in the Social Sciences (LVSS) was to address an area that:

- Offers multiple points of intersection between the social sciences and the statistical and applied mathematical sciences, in particular in order to more extensively introduce applied mathematical modeling in social sciences contexts;
- Is not addressed frequently at other DMS-funded mathematical sciences research institutes;
- Leverages across multiple research centers at the Research Triangle universities, including the Center for Demographic Studies and the Center for Decision Studies at Duke; the Odom Institute for Research in Social Science and Carolina Population Center at University of North Carolina at Chapel Hill (UNC), as well as at organizations such as North Carolina State University (NCSU), RTI International and the SAS Institute.

The organizing theme of the program was *latent variables,* which are widespread in the social sciences. Whether it is intelligence or socioeconomic status, many variables cannot be directly measured. Factor analysis, latent class analysis, structural equation models, error-in-variable models, and item response theory illustrate models that incorporate latent variables. This SAMSI program took a broad look at latent variables and measurement error. Issues of causality, multilevel models, longitudinal data, and categorical variables in latent variable models are examples of the SAMSI topics for this program.

A great success of the program, which was evident from the Kickoff Workshop and the spin-off working groups, is that the program has brought together social scientists, statisticians, biostatisticians, and mathematicians who otherwise would not have gotten together. Members of these disciplines are no different than other disciplines in that their typical interactions are with others from the same discipline. When mixing occurs it is most likely for disciplines that are close in subject matter (e.g., psychology with sociology or statistics with biostatistics). Diffusion of knowledge across disciplines can be a slow process. Furthermore, the latent variable models developed in the social sciences are typically not known in statistics and biostatistics. However, statistics and biostatistics are much more interested in latent variable models today than they were a decade ago.

The program has provided—and through continuing collaborations will continue to provide—a more rapid sharing of knowledge across disciplines than would have occurred without it. The program has clearly generated collaborations across these disciplinary lines that would have not otherwise have occurred.

2 Working Group Activities

2.1 Categorical Variables

Many categorical observed variables in the social sciences are imperfect measures of underlying latent variables. The working group on categorical observed variables was formed to study the statistical issues that emerge when categorical observed variables are part of a model with latent variables or with measurement error. Long-distance participation in this group has been particularly lively. Roland Thomas, a business school professor and statistician of Carleton University in Canada regularly called in and was active in developing the group's research agenda, while Liqung Wang, a statistician, actively participated in E-mail discussions. The group met weekly during the fall and spring semester.

The working group began by studying the relationship between two approaches to models with categorical outcomes: James Hardin's and others' recent advances in models that correct for measurement error in nonlinear models within the GLM framework, and Ken Bollen's two-stage least squares approach to latent variable models.

After contrasting the differences and similarities in these approaches, we began to pursue specific research papers that grew out of this examination of the different approaches. The group is currently drafting a paper entitled "Limited dependent variable models with errors in covariates." James Hardin, Ray Carroll, and colleagues have examined an instrumental variable approach to generalized linear models that corrects for measurement error in covariates. In their work, published in the *Journal of the American Statistical Association* and elsewhere, they describe their technique as "approximately consistent" with the consistency related to the degree of measurement error in the covariates. Several members of the working group derived methods to permit consistent estimation of the coefficient parameter under the same conditions. The paper being written focuses on the theoretical derivation and application to empirical examples. Co-authors include Professors Bollen, Thomas, and Wang. Future planned research may include a paper on Bayesian approaches to the problem of nonlinear models with measurement error; Jane Zavisca, a SAMSI postdoc, and Saki Kinney, a graduate student in Statistics at Duke, have begun a literature review to that end.

2.2 Complex Surveys

This working group worked on a range of problems for analysis of survey data. Since most of the groups' participants were long-distance, and their interests clustered into two categories, the group split early in the program into two subgroups with the following themes: Latent Class Analysis (LCA) of Measurement Error in Surveys, and Weighting and Estimation for Complex Sample Designs. Each subgroup met approximately once a month via teleconference, and also corresponded via a listserv between meetings.

LCA Subgroup: This subgroup is developing latent class models for assessing measurement error in survey responses, with the goal of improving questionnaire design. Many questions in social surveys have categorical, multiple choice responses, making latent class analysis an appropriate technique for estimating measurement error when repeated measures are available. Group members included Paul Biemer, the group leader, Bac Tran of the US Census Bureau, Clyde Tucker of the Bureau of Labor Statistics, Brian Meekins of the Bureau of Labor Stastics, and Jane Zavisca, a postdoc at SAMSI.

The group's primary work to date has concentrated on the problem of rotation group bias in longitudinal surveys, using the empirical example of labor force classification in the Current Population Survey. The census bureau provided the group with access to specialized data that permits fitting a wide range of LCA models for measurement error. Jane Zavisca presented preliminary findings at the American Sociological Association Section on Methodology 2005 Annual Meeting. We plan to finish analysis and submit a paper for peer review this summer. Our current work focuses on the cross-sectional re-interview sample; in the future we will apply Markov LCA models to panel data (for which re-interviews are not available).

Although the unemployment example has been the primary focus of our research, we have also read papers and discussed ongoing work by group members on related topics. For example, we have discussed modeling measurement error when the latent variable of interest is thought to be continuous rather than categorical, with applications to quality of response in surveys of adolescent drug and alcohol use. All group members have broader research agendas for which this type of modeling is appropriate, and future collaboration beyond the SAMSI program is likely.

Weighting and Estimation Subgroup: The complex sampling working group (CSWG) was the other LVSS survey subgroup. The subgroup had approximately 15 active members and was led by Chris Skinner (Southampton). Membership was international and cooperation was established with the Canadian National Program on Complex Data Structures (NPCDS), supported by the Canadian National Science and Engineering Council. NPCDS supported a workshop in Montreal on May 4-6 2005 on "Latent Variable Models and Survey Data for Social and Health Sciences Research" as a contribution to the LVMSS programme. The work of the CSWG subgroup took place between the LVMSS kick-off workshop in September 2004 and the Montreal workshop and preparation for the latter workshop, led by Mary Thompson (Waterloo), was a significant component of the subgroup's work. Since subgroup membership was international with only a few members local to SAMSI, most communication occurred via email and conference call. The subgroup met monthly between October 2004 and April 2005 via conference call to keep abreast of the progress and happenings. Communication via the listserv was ongoing and included sharing research, feedback on research, question and answers, and discussion as well as preparation for the workshop. Several members were working on independent projects dealing with weighting in multi-level modeling. The subgroup discussed the need for a document providing the most current information on weighting in multilevel models and discussed its contents. Chris Skinner (Southampton) and Kim Chantala (UNC) have taken forward preparation of this paper, using analyses of data from an adolescent health survey and plan to submit this paper for publication.

The subgroup members met in person at the Montreal conference. The programme committee consisted of Mary Thompson (Waterloo, chair), Chris Skinner (Southampton), Paul Biemer (North Carolina), Jamie Stafford (Toronto), Milorad Kovacevic (Statistics Canada), Randy Sitter (Simon Fraser), David Bellhouse (Western Ontario), Roland Thomas (Carleton). The conference programme is available on www.crm.umontreal.ca/Latent05. Subgroup members making presentations included Sharon Christ (UNC), Georgia Roberts (Statistics Canada), Adam Carle (Bureau of the Census), Stanislav Kolenikov (UNC), Kim Chantala (UNC) and Chirayath Suchindran (UNC). Bengt Muthen and Timo Asparouhov (UCLA) made presentations about their survey data modeling project, which is implementing complex sample analysis capabilities in the latent variable modeling software MPlus. In addition, A. Skrondal and S. Rabe-Hesketh provided a tutorial on generalized latent variable modeling and there were a number of invited papers, including from Ken Bollen.

2.3 Longitudinal Data

The participants in the working group fell into three categories: 1. MetaMetrics; 2. Chris Kelly (Post-doc, UNC Institute of Aging); 3. Minimal participation by 2 or 3 others. Representatives from MetaMetrics (5-6 people) have made up the bulk of the group and consistently attended the meetings. Chris Kelly has attended many of the meetings but he and Lloyd Edwards have consistently met outside of SAMSI (at UNC) to address his needs.

Initially, an assessment was made regarding the needs and interests of the various participants in the group. As a result, since the inception of the program the working group has covered what amounts to a semester's worth of graduate material in applied longitudinal data analysis using the linear mixed model. Both MetaMetrics and Chris Kelly have made presentations to the group.

As a result of the LVSS Program, a long-term collaborative effort has been identified between MetaMetrics and Lloyd Edwards. MetaMetrics and Edwards are working together to address methodological and applied aspects of longitudinal data analysis of a very large dataset. The objective is to determine the profile of standardized reading scores across time and accurately assess predictors of the reading scores. The development of standardized reading scores, called Lexiles, is company proprietary information. In addition, since the dataset is large, we will research the use of cross-validation methods in longitudinal data analysis using the linear mixed model.

Chris Kelly and Lloyd Edwards are working together on a publication from his doctoral research regarding factors predicting the extent of nursing home regulation in the 50 states of the U.S. Kelly originally attempted to apply ordinary least squares and hierarchical linear modeling techniques to his longitudinal study. After publication rejection, Kelly and Edwards are working to apply appropriate linear mixed model techniques and are having great success.

2.4 Model Uncertainty

The primary goal of the Model Uncertainty Working Group was to develop new statistical methods for accommodating model uncertainty in latent variable models motivated by social science applications. Because structural equation models (SEMs) with latent variables provide a broad and flexible framework for analysis of multivariate data, the working group chose to focus on uncertainty that arises in specifying an SEM as a starting point. The group began by reviewing the literature on model selection and averaging in SEMs. The BIC criteria, which was originally proposed as an approximation to the Bayes factor, was chosen as a good starting point for discussion. There are a series of articles in the literature by Adrian Raftery and others arguing in favor of the BIC.

Research Accomplishments and Ongoing Work

Several members of the working group were interested in exploring improvements to the BIC, which could be calculated using the output produced by standard software for fitting SEMs. This work has lead to several papers:

Bollen, K.A., Ray, S. and Zavisca, J. (2006a). A scaled unit information prior approximation to the Bayes factor. Submitted for publication.

Bollen, K.A., Ray, S. and Zavisca, J. (2006b). A new approximation to the Bayes factor for structural equation model selection. In progress.

Visser, I., Ray, S., Jang, W., Berger, J., Bayarri, S. and Pericchi, L. (2006). Generalization of BIC. In progress.

It has been noted previously that selecting models using the BIC corresponds approximately to using a Bayes factor with an implicit unit information prior. Such a prior may not be a good choice for model selection, because it implies a very stringent penalty on model complexity, which tends to overly-favor smaller models. Motivated by this problem, Bollen, Ray and Zavisca (2006a) consider an approximation to the Bayes factor based on a scaled unit information prior. In ongoing work, Bollen, Ray and Zavisca (2006b) apply different approximations to the Bayes factor to the problem of structural equation model selection and compare the results. In hierarchical models, a vexing problem in deriving model selection criteria is that it is not clear what the penalty for model complexity should be. Specifically, it is not clear what the effective sample size is. Motivated by this problem, Visser et al. (2006) propose a generalization of the BIC, deemed the GBIC. The GBIC is derived by standard Laplace approximation routes, but with the prior not included in the Laplace expansion. Various default priors for the component parameters are considered, with the scale of the prior chosen based on a novel definition of effective sample size, which allows for different effective sample sizes for the different parameters. This is clearly an appealing idea, as some parameters may be weakly identified from the data, while others can be estimated very reliably.

It was agreed by the working group that Bayesian methods provide an appealing framework for accommodating model uncertainty in SEMs, allowing for both model averaging and selection. The working group therefore reviewed the literature on Bayesian SEMs, which was found to be quite limited, motivating a summary article (to be included in the Handbook on Structural Equation Models, to be published by Elsevier):

Palomo, Jesus, David Dunson, and Kenneth A. Bollen. Forthcoming. "Bayesian Structural Equation Modeling" in *Handbook on Structural Equation Models*.

> *Structural equation models (SEMs) with latent variables are routinely used in social science research, and are of increasing importance in biomedical applications. Standard practice in implementing SEMs relies on frequentist methods. This chapter provides a simple and concise description of an alternative Bayesian approach. We provide a brief overview of the literature, describe a Bayesian specification of SEMs, and outline a Gibbs sampling strategy for model fitting. Bayesian inferences are illustrated through an industrialization and democratization case study from the literature. The Bayesian approach has some distinct advantages, due to the availability of samples from the joint posterior distribution of the model parameters and latent variables, that we highlight. These posterior samples provide important information not contained in the measurement and structural parameters. As is illustrated using the case study, this information can often provide valuable insight into structural relationships.*

The Bayesian posterior computation for SEMs relies on Gibbs sampler, with conditionally conjugate priors used for convenience in implementation. However, as in other hierarchical models, it can be unclear how to best choose default priors for the variance components and other parameters. Commonly-used vague inverse-gamma priors lead to poorly-behaved algorithms for posterior computation, and results can be sensitive to arbitrarily-chosen hyperparameters. In simple variance component models, Gelman (2005) proposed a class of half-t priors, which avoid these problems. Palomo and Dunson (2006) generalized these priors to the setting of SEMs and developed an efficient parameter-expanded Gibbs sampling algorithm for posterior computation:

> *"Bayesian Inference and Computational Issues in Structural Equation Modeling"*
> *Jesus Palomo and David Dunson, to be submitted.*

This approach does not require subjective choice of hyperparameters and leads to much more efficient computation. Motivated by the parameter-expansion approach and by the need for efficient algorithms for accommodating model uncertainty in the number of latent factors in a factor analytic model, Dunson (2006) developed a stochastic search factor selection (SSFS) algorithm:

> *"Efficient Bayesian model averaging in factor analysis"*
> *David Dunson, ISDS Discussion Paper2006-03, to be submitted.*

Factor analytic and structural equation models are intimately related to random effects models commonly used for analysis of longitudinal data. In analysis of random effects models, the problem of selecting predictors having random effects while also performing inferences on heterogeneity in the predictor effects is difficult. One issue is that the null hypothesis of no heterogeneity corresponds to a random effect variance of zero, which is a value on the boundary of the parameter space. This boundary problem invalidates typical likelihood ratio test asymptotic results, so new methods are needed. Motivated by this problem, Satkartar Kinney focused on developing a Bayesian approach to account for model uncertainty in linear and logistic random effects models. This work was very successful and has resulted in the paper:

"Fixed and random effects selection in linear and logistic models"
Satkartar Kinney and David Dunson, ISDS Discussion Paper 2006-06, submitted

The proposed method should also be useful for efficient Bayesian analysis of a single linear, probit or logistic mixed effect model, with default priors chosen for the parameters.

In addition to uncertainty in the predictors to be included in the fixed and random effects components of the model, there is also uncertainty in the assumed parametric form for the distribution of the random effects. For this reason, it is very appealing to consider semiparametric Bayesian methods that avoid parametric assumptions on the random effects distribution, while also allowing for selection of predictors having varying coefficients. The following paper addressed this problem using Dirichlet process priors with a variable selection mixture prior incorporated in the base measure:

"Variable selection in nonparametric random effects models"
Bo Cai and David Dunson, ISDS Discussion Paper2005-16, submitted.

Ongoing work focuses on extending these methods to structural equation models, while also developing more efficient approaches, which can be applied quickly to large data sets, such as are often collected in population studies. One motivation is the North Carolina Agricultural Health Study, which collected neurological symptom data for over 30,000 farm workers exposed to pesticides and other factors. Efficiency can potentially be improved by using a combination of accurate approximations and stochastic simulation.

The model uncertainty working group was clearly extremely productive and has stimulated a series of ongoing collaborations. In addition, the lack of good methods for solving important problems in this area is very apparent. For example, it is typically the case that many structural equation models are consistent with sociologic or biologic theory a priori. In fact, the number of plausible models may be in the 1,000s or even more. However, there is no good way to take into account this uncertainty in assessing hypotheses about particular structural relations of interest. To address the gap in the literature and encourage more work in this interesting area, David Dunson will edit a book published by Springer on “Model uncertainty in random effects and latent variable models.” This book was directly motivated by the working group and several of the active participants will contribute chapters.

2.5 Multilevel Models

Over the past year, the working group explored a number of issues with respect to fitting multilevel models with latent variables (also known as hierarchical, random coefficients, or mixed-effects models). Our primary goal was to examine and/or develop models that could allow for random effects when either or both the criterion and predictors were constructs that could not be measured directly and without error. The general structure our group followed was to complete and discuss targeted readings and present illustrative analyses at the working group meetings.

In the fall, we began by considering multilevel models with random coefficients but without latent variables – that is, where the criterion and predictors are all assumed to be perfect (error-free) measures of the constructs they are said to represent. We examined both frequentist (likelihood-based) and Bayesian approaches to fitting these models. Using this as a base, we then extended our consideration to multilevel models with latent variables. We read work on the correspondence between multilevel models and structural equation models, including papers by Bauer (2003) and Skrondal and Rabe-Hesketh (2004). These readings allowed us to contextualize and compare two basic multilevel latent variable models originating in different literatures. Although useful, these models apply only under a limited set of conditions. Specifically, these models are linear in their parameters, require continuous observed variables, and allow only random intercepts or means in addition to the latent variables.

In the spring, we examined ways to expand the model to overcome these limitations. One key extension was to allow for categorical observed variables, as these are quite frequent in social science research. We read and discussed further papers by Kamata (2001), Skrondal & Rabe-Hesketh (2004), Rijmen et al. (2003), and Thissen & Orlando (2001) on the correspondence between item-response theory models, binary factor analysis, and nonlinear mixed models. We were fortunate to have Anders Skrondal and Sophia Rabe-Hesketh come to SAMSI in April to present on their approach to fitting multilevel latent variable models with continuous or categorical observed variables. The second key extension that we examined with these models was the possibility to allow for both random intercepts and slopes; specifically, to have parameters of the measurement model (relating the latent variables to the observed variables) or structural model (relating the latent variables to one another) vary over units. The primary optimization difficulty in allowing both random intercepts and slopes is that one must integrate over more dimensions with a model that is nonlinear in the unknown parameters. A secondary difficulty is the need to have many observations per unit and many units to strongly identify and make possible the estimation of the model. We considered several possible ways of addressing these difficulties, with various strengths and weaknesses. These included the use of quadrature, Markov Chain Monte Carlo, an approximate method based on finite mixture models, and an approximate method that assumes the “random” effects are predetermined functions of other variables.

Two papers that have so far grown out of this group activity are:

- Kamata, A. & Bauer, D. J. (2005). A Note on the Relationship between Factor Analytic and Item Response Theory Models. Under review for Applied Psychological Measurement.
- Kamata, A., Bauer, D. J., & Miyazaki, Y. (2005). Multilevel Measurement Model. To appeared in A. A. O'Connell & D. B. McCoach (Eds.). Multilevel Analysis of Educational Data. Information Age Publishing.

2.6 Social Networks

This working group is an outgrowth of the NSF-funded NISS project *Dynamics for Social Networks Processes: Comparing Statistical Models with Intelligent Agents*. The thrust of the project is to reconcile two methods for modeling change in social networks over time--*p** models and intelligent agent models. The latter family has received much attention from social scientists but little from mathematicians and almost none from statisticians, and so constitutes a promising and important opportunity for collaboration. The group pursued three complementary approaches to modeling dynamics of social networks:

- Using stochastic differential equations to describe evolution of participants characteristics and relationships (“edges”): H.T. Banks, Karr, Nguyen, Samuels, leading to the paper

  > Banks, H.T., A.F. Karr, H.K. Nguyen, and J.R. Samuels, Jr. (2005). Sensitivity to noise Variance in a Social Network Dynamics Model. *Quarterly of Applied Mathematics* (to appear).

- A mathematical programming formulation that maximizes “affinity:” Medhin, Hong.
- A discrete-time formulation with (initially) fixed relationships and participants that move with a “social space” of (possibly latent) characteristics: D. Banks, Chu.

Principal senior participants are David Banks (Statistics, Duke), H. T. Banks (Mathematics, NCSU), Kenneth Bollen (Sociology, UNC), Kathleen Carley (Computer Science, Carnegie Mellon), Alan Karr (NISS) and Negash Medhin (Mathematics, NCSU). SAMSI/CRSC postdoctoral Nguyen Hoan has also participated, as have graduate fellows John Hipp (Sociology, UNC) and Johnny Samuels (Mathematics, NCSU) and graduate associates and Jen-hwa Chu (Statistics, Duke), Chung-Chien Hong (Mathematics, NCSU) and Eric Vance (Statistics, Duke).

3 Personnel

3.1 Faculty Releases and Associates

Faculty releases for the program were Kenneth Bollen (Sociology, UNC), Lloyd Edwards (Biostatistics, UNC), Subhashsis Ghosal (Statistics, NCSU) and Negash Medhin (Mathematics, NCSU).

Dan Bauer (Sociology, UNC), Paul Biemer (RTI International) and David Dunson (NIEHS) led working groups as faculty associates. David Banks (Statistics, Duke), H. T. Banks (Mathematics, NCSU) and Alan Karr (NISS) participated in the social networks working group.

3.2 Research Visitors

Long-term research visitors to the LVSS program included Maria Jesus Bayarri, University of Valencia; Aki Kamata, Florida State University; Katja Ickstadt, University of Dortmund; Xiadong Lin, University of Cincinnati; Ingmar Visser, University of Amsterdam.

Targeted experts, here for shorter visits, included Yasuo Miyazaki, Virginia Tech; Sophia Rabe-Hesketh, Berkeley; Anders Skrondal, London School of Economics.

3.3 Postdoctorals

Jane Zavisca has been active in most of the six working groups in the LVSS program. In the latent class subgroup of the survey working group, Jane has worked closely with Paul Biemer, a research statistician at the Odum Institute for Research in Social Science and RTI. She presented a coauthored a paper on their work on a latent class analysis of unemployment statistics at the spring meeting of the Methodology Section of the American Sociological Association in Chapel Hill, North Carolina. Jane also is working closely with the Model Uncertainty working group. Here, too, she was coauthor on a paper presented to the Methodology Section of the American Sociological Association. The paper was on approximations to the Bayes factor in Structural Equation Models. Both the latent class unemployment and the Bayes factor paper will be revised and sent out for review for publication in refereed journals. Jane also has been an active participant in the Multilevel and Structural Equation Modeling working group and the Categorical Observed Variable and Latent Variables working group. In addition, to contributing to the discussions and preparing results for the group, Jane helped with some of the web site maintenance for several of these groups.

She was also a major contributor to the February 18-19, 2005 Undergraduate Workshop, presenting material from the LVSS program.

Jesus Palomo and **Surajit Ray** were major contributors to the Model Uncertainty working group.

3.4 Graduate Fellows and Associates

John Hipp (Sociology, UNC) played an active role in several of the working groups. He was a regular participant in the Multilevel and Structural Equation Model group. He regularly attended and participated in the group discussions. John also was a member of the Categorical Observed Variables in Latent Variable Models group. For this group, he provided empirical and simulation examples that enabled us to explore several of the issues that emerged in our discussion. He also used these to demonstrate some of the new analytical results that we developed. John also participated in the social network working group. He is knowledgeable about social network data, theory, and examples and this knowledge was useful to the group. Hipp's dissertation to be completed this spring makes use of latent variable techniques in his analysis of neighborhood satisfaction using survey data collected by the U.S. Census. Finally, John Hipp helped with some of the logistics of making the working groups operate smoothly.

John Samuels (Mathematics, NCSU) has participated in the Social Networks working group that meets every Thursday at SAMSI. He has pursued specific research with Alan Karr, Hoan Nguyen (a SAMSI postdoctoral Fellow) and H.T. Banks. During the year they have developed a model for social dynamics (characteristics associated with a number of agents) of buddy/clique formation. The characteristics (on a continuous time, continuous value scale from -10 to +10) are assigned to each agent. A nonlinear model for degree of connectivity is coupled to a nonlinear stochastic differential equation for the evolution of the characteristics in the agent population. The resulting system of SDE is then solved by a classical fourth-order Runge-Kutta discretization procedure. Samuels has contributed substantially to the research and methodology and will be a co-author on a forthcoming publication this summer. He also played a major leadership role in the SAMSI Undergraduate Workshop held May 30-June 3, 2005.

Jen-hwa Chu (Statistics, Duke) has been involved in building agent-based models of social network dynamics. These models incorporate the latent-variable space approach described by Hoff, Raftery and Handcock (2002), as well as covariate information such as gender and memory of past relationships. The intent is to build rule-sets such that the dynamics of agent behavior mirror the dynamical models being from two other perspectives by other teams in the working group. The comparison of the models is based upon summary statistics from repeated runs, such as the mean and standard deviation of the number of persistent cliques, the first three moments of the in-degrees, the mean and standard deviation of the number of triad completions, and so forth.

Jiezhun Gu (Statistics, NCSU) has also been involved with the program. She has attended all of the weekly meetings, and has expressed interest in working on the asymptotic theory associated with model uncertainty. Subhashis Ghosal is also interested in working on this problem.

Satkartar Kinney (Statistics, Duke) has also made substantial progress on model uncertainty. She has focused on the problem of selecting fixed and random effects in logistic mixed effects models. A number of methods have been proposed in the literature for subset selection in regression, but little has been done for the challenging problem of selecting predictors that vary in their effects for different individuals and hence have associated variance components. Satkartar initially adapted an approach proposed by Chen and Dunson (2003) for the linear mixed model to the probit case. She then further modified this approach to logistic regression using a data augmentation scheme, with slice sampling used in the implementation. Ongoing work focuses on parameter expansion methods, which lead to more efficient computation and improved priors for the variance components.

Eric Vance (Statistics, Duke) has been studying social network behavior in elephant herds using data collected by ethologists. He has fit the Hoff, Raftery, and Handcock (2002) latent variable version of the dyadic p* model and found that group dynamics change between the wet and dry season, and that genetic relatedness and the social hierarchy play a large role in elephant networks. The inferences are Bayesian, and use Markov chain Monte Carlo to find the posterior distributions of each of these effects. He has written a paper on the social network methodology and it is submitted to the Journal of Organizational Computation.

Chien-Chung Wong (Mathematics, NCSU), working with Medhin, developed a model where each actor is endowed with a set of dynamic personal attributes, values, and preferences, and a set of statistical information on each of the other actors in the social group. If the social network consists of N actors we construct an NxN matrix of zeros and ones, called sociomatrix. If actor i is friendly toward actor j, then the ij-th entry of the matrix will be one, otherwise zero. The diagonal entries of the matrix are set to zero. If the ij-th entry of the matrix is 1, then we say there is a link from i to j. We have developed a model where the ij-th link depends on the maximum of a payoff of an appropriately constructed nonlinear programming problem involving the attributes and values of the actors in the social group. The model can be modified to handle social status as well as general network dependence structure. For example, the link from i to j may not be completely independent of the link from i to k, and/or from j to k.

The model also incorporates migration and preferred attributes. The approach developed captures the ideas of the well established P1 model introduced by Holland and Leinhart, and the more recent extension, P2 model, due to van Duijn, Snijders, and Zijlstra. In particular the model developed reflects reciprocity, and attributes and values of actors i and j play a role in determining whether or not there is linkage between these actors.

4. Workshops

4.1 Opening Workshop

The Opening Workshop, held on September 11-14, 2004, was among the largest and most diverse to date at SAMSI. There were more than 100 attendees.

4.2 Causality Mini-Workshop

A Mini-Workshop on Causality was held on March 29, 2005, with more than 80 attendees.

4.3 GLLAMM Mini-Workshop

Anders Skrondal and Sophia Rabe-Hesketh came to NISS for three afternoons workshop in April 2005. They provided an overview of the GLLAMM model that they have developed for multilevel latent variable, structural equation models. The topic was of interest to several of the working groups and approximately 15 to 20 attended this workshop.

4.4 Yearend Summaries of Working Group Activities

On May 19, 2005 we held an all day conference where each working group presented a summary of the work that they had completed during the academic year. This work is more fully described above under each working group. Approximately 40 to 50 participants attended this meeting.

4.5 Latent Variables in the Social Sciences Transition Workshop

On November 10 and 11, 2005, we held another workshop that grew out of the Latent Variables in the Social Sciences SAMSI program. This provided an update on the projects that were started during the previous year. In addition to a number of papers presented by participants in last year’s programs, we had Peter Bentler (Psychology and Statistics, UCLA) and Leo Goodman (Sociology and Statistics, UC, Berkeley) as invited speakers. Over 100 participants from North Carolina and the rest of the country attended the workshop. It was a great success.

5 Course

Kenneth Bollen (Faculty Fellow) and Jane Zavisca (postdoctoral) presented a seminar course entitled *An Overview of Latent Variable Models in the Social Sciences* during the fall semester of 2004-05. The course provided an overview of latent variable models that are common in the social sciences. It was designed to introduce students to a variety of statistical models that make use of latent or unobserved variables, including factor analysis, latent trait and latent class models, and structural equation models. The instructors and guest speakers lectured on the various models and take student questions, followed by class discussion and/or student presentations. Technical and substantive readings will further explain the models and provide examples of concrete applications.

In addition to attending lectures and reading background materials, students taking the course for credit applied one of the types of models covered in the course to a data set of their choice, and made a 15 minute presentation on their findings to the class.

6 Publications

Completed publications and reports:

- Banks, H.T., Karr, A.F., Nguyen, H. K. and Samuels, J.R.,Jr. (2005). Sensitivity to Noise Variance in a Social Network Dynamics Model. *Quarterly of Applied Mathematics* (to appear).

- Bollen, K.A., Ray, S. and Zavisca, J. (2006a). A Scaled Unit Information Prior Approximation to the Bayes Factor.   Submitted for publication.

- Cai, B., and Dunson, D. (2005). Variable Selection in Nonparametric Random Effects Model. ISDS Discussion Paper 2005-16; submitted for publication.

- Dunson, D. B., Palomo, J. and Bollen, K (2005). Bayesian Structural Equation Modeling. To appear in S.-Y. Lee*, Handbook on Structural Equation Models* Elsevier.

- Hipp, J. (2005). Neighborhood Networks of Social Distance: Do They Predict Neighborhood Satisfaction?" Presented at the International Sunbelt Social Network Meeting, Redondo Beach, CA.

- Kamata, A. and Bauer, D. J. (2005). A Note on the Relationship Between Factor Analytic and Item Response Theory Models. Submitted to *Psychometrika*.

- Kamata, A., Bauer, D. J., and Miyazaki, Y. (2005). Multilevel Measurement Model. To appear in A. A. O'Connell and D. B. McCoach, eds., *Multilevel Analysis of Educational Data*. Information Age Publishing.

- Palomo, J., Dunson, D. and Bollen, K. A. (2005). Bayesian Structural Equation Modeling. To appear in S.-Y. Lee, ed., *Handbook on Structural Equation Models.* Elsevier.

- Visser, I. (2005). DEPMIX: An R Package for Fitting Mixtures of (Latent) Markov Models on Multivariate Mixed Time Series Data. Package and manual with illustrative examples available for download at www.r-project.org.

- Zavisca, J. (2005). Does Money Buy Happiness in Unhappy Russia? Submitted to *American Journal of Sociology*.

Reports in preparation:

- Biemer, P. and Zavisca, J. (2006). Measurement Error in BLS Unemployment Measures.

- Bollen, K.A. and Kolenikov, S. (2006). A Specification Test for Heywood Cases in Latent Variable Models.

- Bollen, K. A., Ray, S., and Zavisca, J. (2006b). Bayes Factors in Structural Equation Models (SEMS): Schwarz's BIC and Other Approximations.

- Bollen, K. A., Thomas, R., Wang, L., and Hipp, J. (2005). Limited Dependent Variable Models with Covariate Measurement Error: A Consistent Instrumental Variable Estimator.

- Dunson, D. B., ed. (2006). *Model Uncertainty in Random Effects and Latent Variable Models.* Springer-Verlag.

- Dunson, D. B. (2006). Efficient Bayesian Model Averaging in Factor Analysis. ISDS Discussion Paper 2006-03.

- Dunson, D.B., Palomo, J., and Zavisca, J. (2005). Bayesian Model Selection and Averaging in Structural Equation Models. To be submitted to *Psychometrika*.

- Kelly, C. M., Leibig, P. S., and Edwards, L. J. (2005). Factors Predicting the Extent of Nursing Home Regulatory Activity in the 50 States.

- Kinney, S., and Dunson, D. B. (2006). Bayesian Fixed and Random Effects Selection For Binary Response Models. ISDS Discussion Paper 2006-06.

- Palomo, J., and Dunson, D. B. (2006). Bayesian Inference and Computational Issues in Structural Equation Modeling.

- Skinner, C. and Chantala, K. (2006). Use of Weights in Multilevel Modeling.

- Visser, I., Ray, S., Jang, W., Berger, J., Bayarri, S., and Pericchi, L. (2006). Generalization of BIC.