Large Sparse Data and Algebraic Statistics: Is There a Connection?

Stephen E. Fienberg

Department of Statistics, Machine Learning Department, Cylab, i-Lab
Carnegie Mellon University

Special Session on Advances in Algebraic Statistics
AMS 2010 Spring Southeastern Sectional Meeting Lexington, KY
March 27-28, 2010

Introduction

- Many of the most active areas of statistical research involve large sparse data problems where the number of variables and/or parameters is large, especially relative to the number of independent observations.
- Standard statistical theory for estimation and results related to asymptotic behavior often fail in such settings.
- The computational tools associated with algebraic statistics are useful often only for low-dimensional problems, e.g., involving a small number of parameters.
- In this presentation I describe how algebraic statistical and the related computational tools can nonetheless provide important insights of value in large sparse contingency table and network settings.

Outline for Remainder of My Talk

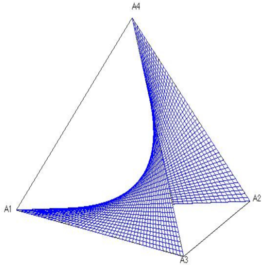
- Five examples and the challenges they have and continue to pose for algebraic statistics:
  1. Example 1—The National Halothane Study
  2. Example 2—The National Long Term Care Survey
  3. Example 3—Monks in a Monastery
  4. Example 4—MIPS Curated PPI in Yeast
  5. Example 5—The Framingham Obesity Study
- Algebraic statistics results and open problems arising from contingency table and network settings.

Example 1—The National Halothane Study

- 50,000 hospital records examined.
- 17,000 deaths arrayed in the form of a very large sparse multi-way contingency table:
  - 34 hospitals
  - 5 anesthetics
  - 5 years
  - 2 genders
  - 5 age groups
  - 7 risk levels
  - type of operation
- Sample of 25 cases per hospital to estimate the denominator, making up the residual 33,000 cases.
- $34 \times 5 \times 5 \times 2 \times 5 \times 7 \times ? = 60,500 \times ?$

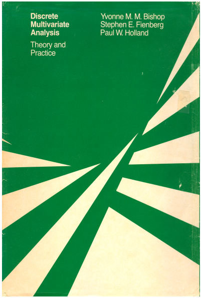
Example 1—Log-linear Models

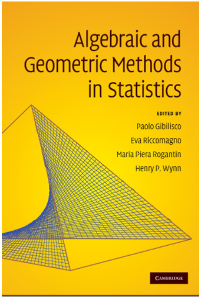
- Work on the Halothane study led to the development of log-linear model theory.
  - Major issue of when MLEs exist for large sparse tables.
- Also geometric representations, especially for 2×2 tables.



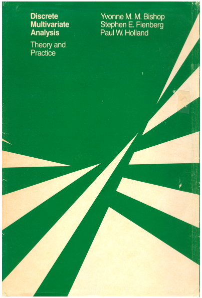
- "Surface of Independence" = *Segre* Variety
- Much later we had:
  - Markov bases for conditional distributions given margins.
  - Representation of log-linear model parameters in terms of polynomial maps.

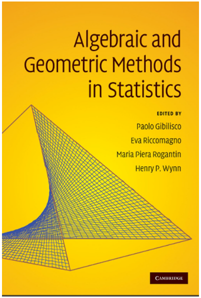
- A tale of two book covers:





- A tale of two book covers:

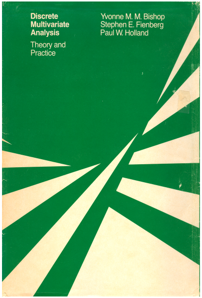


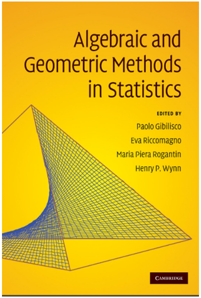


- Computational tools helped with the original Halothane Study problem of existence of MLEs (Ericksen et al. 2006), but only for relatively low dimensional problems.

Algebraic Statistics and Log-linear Models

- A tale of two book covers:





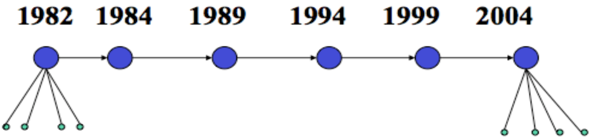
- Computational tools helped with the original Halothane Study problem of existence of MLEs (Ericksen et al. 2006), but only for relatively low dimensional problems.
- Full solution came in Rinaldo's thesis linking algebraic statistics to statistical theory for discrete exponential families.

Example 2—The National Long Term Care Survey

- Longitudinal survey of people aged 65+
- Assess chronic disability
- 6 waves: 1982, 1984, 1989, 1994, 1999, 2004
- Measures ADLs and IADLs:
  - Activities of daily living (ADL): Basic self-care (eating, bathing, etc.)—6 binary measures.
  - Instrumental Activities of Daily Living (IADL): Related to independent living within a community (preparing meals, maintaining finances, etc.)—10 binary measures.
- Each individual that enters the survey is reinterviewed in all subsequent waves until death.
- Approx. $20k$ individuals per wave. $45,009$ unique individuals sampled in all six waves together. Each wave incorporates $\approx 5k$ new subjects to replace those who have died.

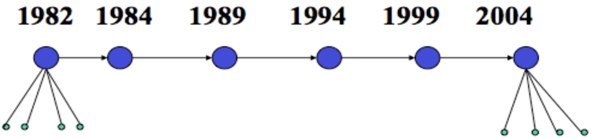
Longitudinal Modeling of NLTCS—Overview

- Sequential measurements on the same individuals allow to assess *individual* disability trajectories over time.



Longitudinal Modeling of NLTCS—Overview

- Sequential measurements on the same individuals allow to assess *individual* disability trajectories over time.

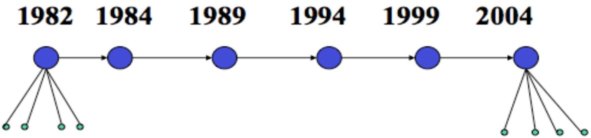


Specifically, we want to

- Understand evolution over time:

Longitudinal Modeling of NLTCS—Overview

- Sequential measurements on the same individuals allow to assess *individual* disability trajectories over time.

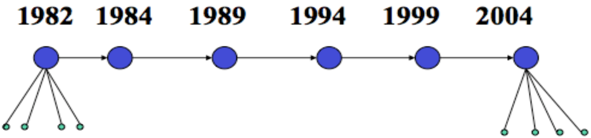


Specifically, we want to

- Understand evolution over time:
  - Individuals

Longitudinal Modeling of NLTCS—Overview

- Sequential measurements on the same individuals allow to assess *individual* disability trajectories over time.

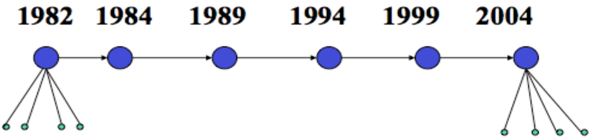


Specifically, we want to

- Understand evolution over time:
  - Individuals
  - Population

Longitudinal Modeling of NLTCS—Overview

- Sequential measurements on the same individuals allow to assess *individual* disability trajectories over time.

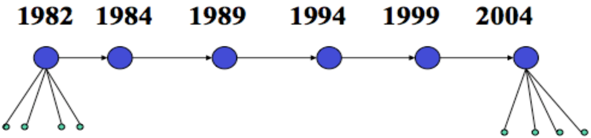


Specifically, we want to

- Understand evolution over time:
  - Individuals
  - Population
- Identify ‘typical’ evolutions over time

Longitudinal Modeling of NLTCS—Overview

- Sequential measurements on the same individuals allow to assess *individual* disability trajectories over time.

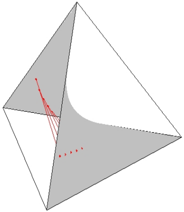


Specifically, we want to

- Understand evolution over time:
  - Individuals
  - Population
- Identify ‘typical’ evolutions over time
- Account for and understand individual variability

Two Types of Models for NLTCS Data

- Latent class models (naive Bayes mixture):
  - Works at the population level.
  - For cross-section or single point in time, see algebraic characterization of Fienberg et. al. (2010).



    - Identification issue clarified by algebraic statistics.
    - Multi-modality of likelihood function.
    - Related to Sturmfels' 100 Swiss Franc problem.
  - For full time-varying latent class model we need a state space structure for the latent classes.
- Mixed-membership models:
  - Works at individual level.

Longitudinal Trajectory Mixed-Membership Model

- Assume the existence of K "ideal classes" or "extreme profiles"
- Assign each individual a *Membership Vector*:

$$g_i = (g_{i1}, g_{i2}, ..., g_{iK})$$

with $g_{ik} > 0$ and $\sum_{k=1}^{K} g_{ik} = 1$ ($g_i \in \Delta_{K-1}$).

- For the "ideal" individuals, specify the marginal distribution of response j, at measurement time t, as a function of some time-dependent covariates.

$$\Pr\left(Y_{ijt} = y_{ijt} \mid g_{ik} = 1, X_i, \theta\right) \quad = \quad f_{\theta_{j|k}}\left(y_{ijt} \mid X_{it}\right)$$

Longitudinal Trajectory Mixed-Membership Model (2)

- Mixed Membership: For a generic individual i, we model

$$\Pr\left(Y_{ijt} = y_{ijt} \,|g_i, X_i, \theta\right) \quad = \quad \sum_{k=1}^{K} g_{ik} f_{\theta_{j|k}}(y_{ijt}|X_{it})$$

- Assuming conditional independence,

$$\Pr\left(Y_i = y_i \,|g_i, X_i, \theta\right) = \prod_{j=1}^{J} \prod_{t=1}^{N_i} \sum_{k=1}^{K} g_{ik} f_{\theta_{j|k}}(y_{ijt}|X_{it})$$

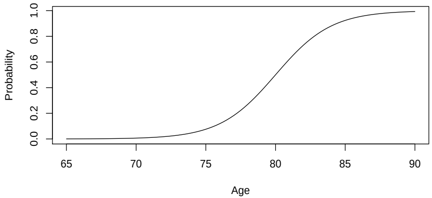
- Assume that the membership vectors are an iid sample from a common distribution (e.g., Dirichlet) with support on the $K-1$ dimensional unit simplex (Δ_{K-1}):

$$g_i|\alpha\text{ı} \overset{iid}{\sim} \text{Dirichlet}(\alpha_0 \times \xi)$$

with $\alpha_0 > 0$ and $\xi = (\xi_1, \xi_2, ..., \xi_K) \in \Delta_{K-1}$.

Basic Model—Extreme Profile Trajectories

- For each extreme profile ($g_k = 1$) specify trajectories of probability of disability in ADLs as a monotone function of Age:



$$\begin{aligned} y_{ijt} &\sim \text{Bernoulli}\left[\lambda_{j|k}(Age_{it})\right] \\ \lambda_{j|k}(X_{it}) &= \text{logit}^{-1}\left[\beta_{0j|k} + \beta_{1j|k} \times Age_{it}\right] \end{aligned}$$

The Mixed-Membership Challenge

- Manrique and Fienberg (2010) use MCMC methods to compute full posterior distribution for this model.

The Mixed-Membership Challenge

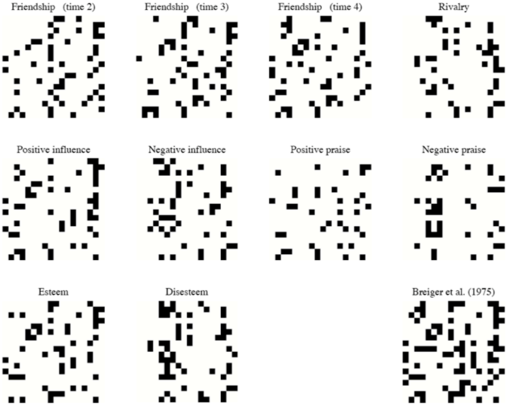
- Manrique and Fienberg (2010) use MCMC methods to compute full posterior distribution for this model.
- Can we exploit hierarchical structure of mixed-membership models to get algebraic statistics characterization?

The Mixed-Membership Challenge

- Manrique and Fienberg (2010) use MCMC methods to compute full posterior distribution for this model.
- Can we exploit hierarchical structure of mixed-membership models to get algebraic statistics characterization?
- Can we relate such a characterization to MCMC methodology?

Example 3: Monks in a Monastery

- 18 novices observed over two years.
- Network data gather at 4 time points; and on multiple relationships.



- See analyses in Airoldi, et al. (2008) *JMLR*.

Holland and Leinhardt's p_1 model

- n nodes, random occurrence of directed edges.

Holland and Leinhardt's p_1 model

- n nodes, random occurrence of directed edges.
- Describe the probability of an edge occurring between nodes i and j:

Holland and Leinhardt's p_1 model

- n nodes, random occurrence of directed edges.
- Describe the probability of an edge occurring between nodes i and j:

$$\log P_{ij}(0,0) = \lambda_{ij}$$

$$\log P_{ij}(1,0) = \lambda_{ij} + \alpha_i + \beta_j + \theta$$

$$\log P_{ij}(0,1) = \lambda_{ij} + \alpha_j + \beta_i + \theta$$

$$\log P_{ij}(1,1) = \lambda_{ij} + \alpha_i + \beta_j + \alpha_j + \beta_i + 2\theta + \rho_{ij}$$

Holland and Leinhardt's p_1 model

- n nodes, random occurrence of directed edges.
- Describe the probability of an edge occurring between nodes i and j:

$$\log P_{ij}(0,0) = \lambda_{ij}$$

$$\log P_{ij}(1,0) = \lambda_{ij} + \alpha_i + \beta_j + \theta$$

$$\log P_{ij}(0,1) = \lambda_{ij} + \alpha_j + \beta_i + \theta$$

$$\log P_{ij}(1,1) = \lambda_{ij} + \alpha_i + \beta_j + \alpha_j + \beta_i + 2\theta + \rho_{ij}$$

- 3 common forms:

  $\rho_{ij} = 0$ (no reciprocal effect)
  $\rho_{ij} = \rho$ (constant reciprocation factor)
  $\rho_{ij} = \rho + \rho_i + \rho_j$ (edge-dependent reciprocation)

Estimation for p_1

- The likelihood function for the p_1 model is clearly of exponential family form.
- For the constant reciprocation version, we have

$$\log p_1(x) \propto x_{++}\theta + \sum_i x_{i+}\alpha_i + \sum_j x_{+j}\beta_j + \sum_{ij} x_{ij}x_{ji}\rho \quad (1)$$

- Holland-Leinhardt explored goodness of fit of model empirically by comparing $\rho_{ij} = 0$ vs. $\rho_{ij} = \rho$.
  - The problem is that standard asymptotics (normality and chi-squared goodness of fit tests) aren't applicable as the number of parameters increases with the number of nodes.
- Fienberg and Wasserman used the edge-dependent reciprocation model to test $\rho_{ij} = \rho$.
- See Goldenberg et al. (2010) review of these and related models.

Algebraic Statistics Results

- Work done in collaboration with Sonja Petrović and Alessandro Rinaldo:
  - Computation of Markov basis elements for $n = 3, 4, 5$.
  - General results follow from computations leading to:

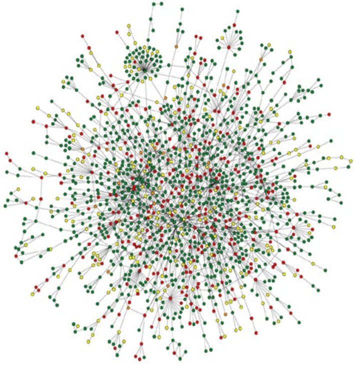
Conjecture

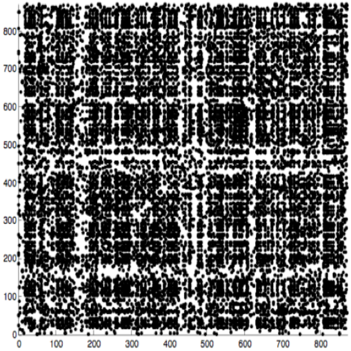
We can obtain minimal Markov (Gröbner) bases for the p_1 models from Markov (Gröbner) bases of $I_{\mathcal{A}_n}$ (the toric ideal of the edge subring of the graph G_n) by repeated lifting and overlapping of the binomials in the minimal Markov bases of various $(n-1)$-node subnetworks.

- How to use results for (1) existence of MLEs and (2) to assess fit of p_1 to large-scale network settings?

Example 4—MIPS-Curated PPI in Yeast

- 871 proteins participate in 15 high-level functions
- Graph and adjacency matrix representations





- Airoldi et al. (2008). *JMLR*.

Example 5: The Framingham Obesity Study

- Framingham Study originated in 1940s and focused on heart disease.
- Offspring cohort of $n_0 = 5124$ individuals measured beginning in 1971 for $T = 7$ epochs centered at 1971, 1981, 1985, 1989, 1992, 1997, 1999.
- Link information on family members and one close friend. Total number of individuals on whom we have obesity measures is $n = 12,067$.
- Christakis and Fowler, *NEJM*, July 2007.

Example 5: The Framingham Obesity Study

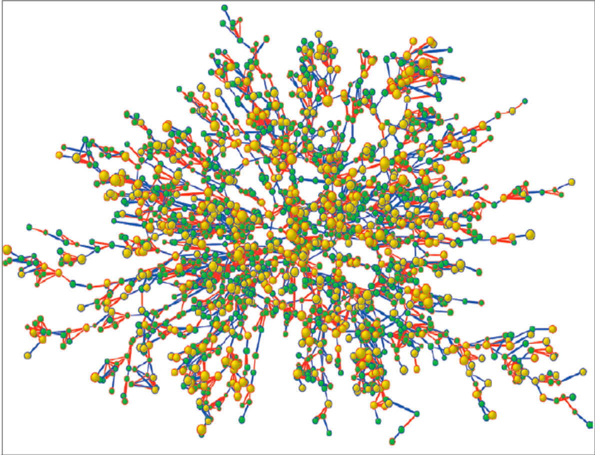


Figure 1. **Largest Connected Subcomponent of the Social Network in the Framingham Heart Study in the Year 2000.**

Each circle (node) represents one person in the data set. There are 2200 persons in this subcomponent of the social network. Circles with red borders denote women, and circles with blue borders denote men. The size of each circle is proportional to the person's body-mass index. The interior color of the circles indicates the person's obesity status: yellow denotes an obese person (body-mass index, ≥30) and green denotes a nonobese person. The colors of the ties between the nodes indicate the relationship between them: purple denotes a friendship or marital tie and orange denotes a familial tie.

Network Model Challenges

- How to use algebraic statistics results for (1) existence of MLEs and (2) to assess fit of p1 to large-scale network settings?

Network Model Challenges

- How to use algebraic statistics results for (1) existence of MLEs and (2) to assess fit of p1 to large-scale network settings?
- Linking algebraic statistics for loglinear models to results for p_1.

Network Model Challenges

- How to use algebraic statistics results for (1) existence of MLEs and (2) to assess fit of p1 to large-scale network settings?
- Linking algebraic statistics for loglinear models to results for p_1.
- Extending results from p_1 to *Exponential Random Graph Models*.

Network Model Challenges

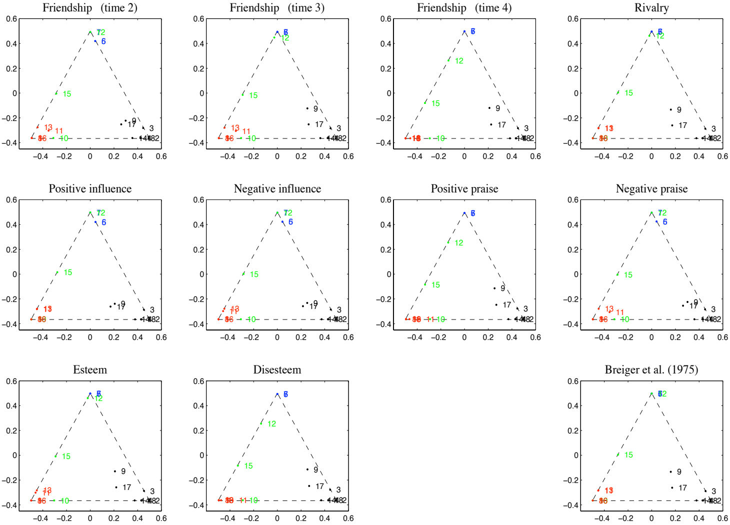
- How to use algebraic statistics results for (1) existence of MLEs and (2) to assess fit of p1 to large-scale network settings?
- Linking algebraic statistics for loglinear models to results for p_1.
- Extending results from p_1 to *Exponential Random Graph Models*.
- Algebraic statistics for mixed-membership stochastic blockmodels.

Network Model Challenges

- How to use algebraic statistics results for (1) existence of MLEs and (2) to assess fit of p1 to large-scale network settings?
- Linking algebraic statistics for loglinear models to results for p_1.
- Extending results from p_1 to *Exponential Random Graph Models*.
- Algebraic statistics for mixed-membership stochastic blockmodels.
- Algebraic statistics characterization of dynamic network models.

MMSB Model for Monk Data

$K = 3$ blocks and extreme profiles



Moral of My Story

- My examples come from contingency table settings and an array of problems involving network structures.

Moral of My Story

- My examples come from contingency table settings and an array of problems involving network structures.
  - They all involve large sparse data problems where the number of variables and/or parameters is large, especially relative to the number of independent observations.

Moral of My Story

- My examples come from contingency table settings and an array of problems involving network structures.
  - They all involve large sparse data problems where the number of variables and/or parameters is large, especially relative to the number of independent observations.
- The computational tools associated with algebraic statistics are often only useful for low-dimensional problems, e.g., involving a small number of parameters.

Moral of My Story

- My examples come from contingency table settings and an array of problems involving network structures.
  - They all involve large sparse data problems where the number of variables and/or parameters is large, especially relative to the number of independent observations.
- The computational tools associated with algebraic statistics are often only useful for low-dimensional problems, e.g., involving a small number of parameters.
- I have described how algebraic statistical and the related computational tools can nonetheless provide important insights of value in large sparse settings.

Moral of My Story

- My examples come from contingency table settings and an array of problems involving network structures.
  - They all involve large sparse data problems where the number of variables and/or parameters is large, especially relative to the number of independent observations.
- The computational tools associated with algebraic statistics are often only useful for low-dimensional problems, e.g., involving a small number of parameters.
- I have described how algebraic statistical and the related computational tools can nonetheless provide important insights of value in large sparse settings.
- There remain many challenges for algebraic statistics in these contingency table and network modeling settings.

Bibliography—Algebraic Statistics Papers

Fienberg, S. E., Hersh, P., Rinaldo, A., and Zhou, Y. (2010) Maximum likelihood estimation in latent class models for contingency table data. In P. Gibilisco, et.al. eds., *Algebraic and Geometric Methods in Statistics*, Cambridge University Press, 31–66.

Fienberg, S. E., Petrović, S., and Rinaldo, A. (2010) Algebraic statistics for p_1 random graph models: Markov bases and their uses. In S. Sinharay and N. J. Dorans, editors, Papers in Honor of Paul W. Holland. Educational Testing Service.

Petrović, S., Rinaldo, A., and Fienberg, S. E. (2010) Algebraic statistics for a directed random graph model with reciprocation," *Proceedings of the Conference on Algebraic Methods in Statistics and Probability,* Contemporary Mathematics Series, AMS.

Rinaldo, A. Fienberg, S. E., and Zhou, Y. (2009) On the geometry of discrete exponential families with application to exponential random graph models. *Electronic Journal of Statistics*, **3**, 446–484.

Slavkovic, A. and Fienberg, S. E. (2010) Algebraic geometry of 2×2 contingency tables. In P. Gibilisco, et.al. eds., *Algebraic and Geometric Methods in Statistics*, Cambridge University Press, 67–85.

Bibliography—Statistics References

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

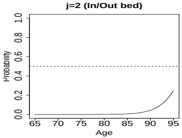
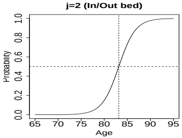
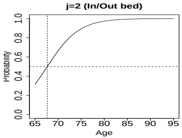
Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975) *Discrete Multivariate Analysis: Theory and Practice.* MIT Press. Reprinted by Springer (2007).

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi E. M. (2010) A Survey of Statistical Network Models, (*Foundations and Trends in Machine Learning*, **2** (2), 129–233.

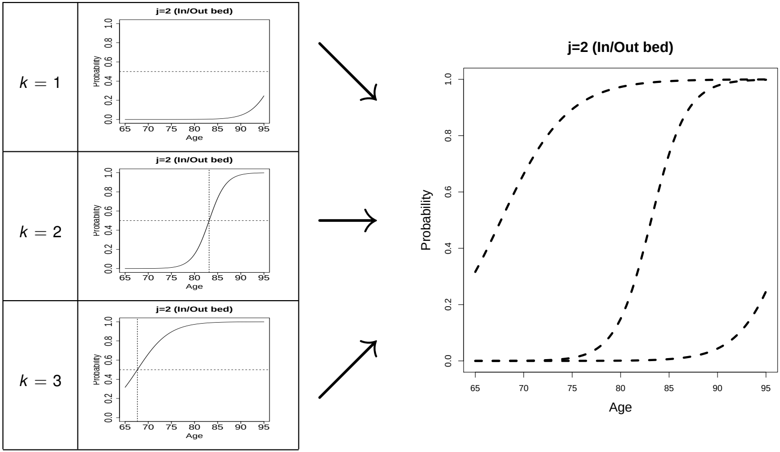
Manrique-Vallier, D. and Fienberg, S. E. (2010) Longitudinal mixed-membership models for survey data on disability. In *Longitudinal Surveys: from Design to Analysis: Proceedings of XXV International Methodology Symposium, 2009*, (2010), Statistics Canada, to appear.

The End

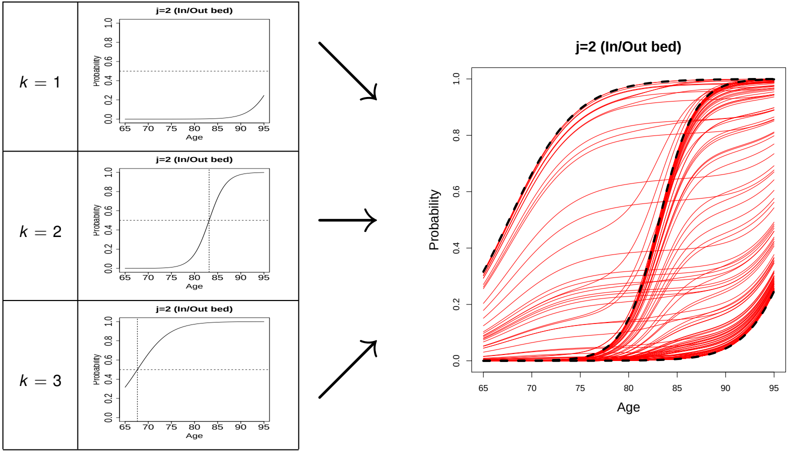
Test Computations—From profiles to Individuals

$k = 1$	
$k = 2$	
$k = 3$	

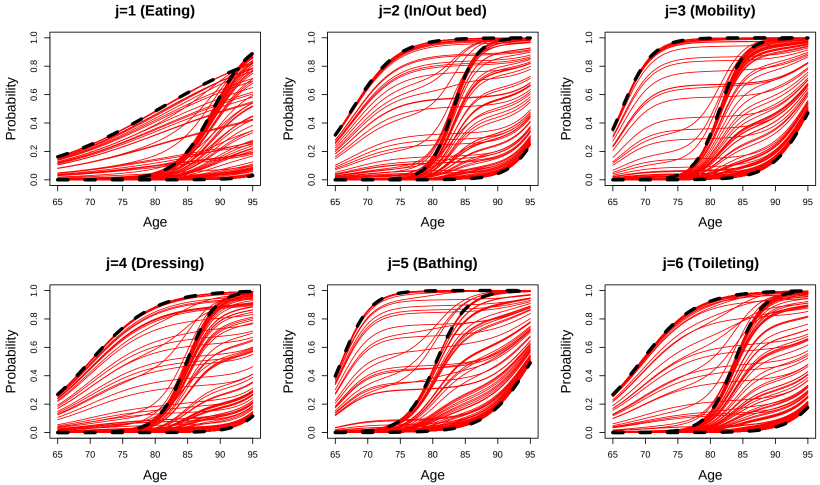
Test Computations—From profiles to Individuals



Test Computations—From profiles to Individuals



Test Computations—Individual Trajectories



Monks in a Monastery

ID	faction	name	order monk left monastery
1	2	Ambrose	9
2	1	Boniface	15
3	1	Mark	7
4	1	Winfrid	12
5	3	Elias	17
6	3	Basil	3
7	3	Simplicius	18
8	2	Berthold	6
9	1	John Bosco	1
10	4	Victor	8
11	2	Bonaventure	5
12	4	Amand	13
13	2	Louis	11
14	1	Albert	16
15	4	Ramuald	10
16	2	Peter	4
17	1	Gregory	2
18	1	Hugh	14