8

Interpretability Constraints and Trade-offs in Using Mixed Membership Models

Burton H. Singer

Emerging Pathogens Institute, University of Florida, Gainesville, FL 32610, USA

Marcia C. Castro

Harvard School of Public Health, Boston, MA 02115, USA

CONTENTS

Although shared membership of individuals in two or more categories of a classification scheme is a distinguishing feature of the family of mixed membership models, relatively few analyses using these models pay much attention to this special feature. Most published analyses to-date focus on identifying and interpreting the extreme, or ideal, types consistent with a given body of data, thereby in effect using mixed membership models as crisp clustering techniques. Getting into the domain of shared membership quickly places the investigator in a difficult position, as standard estimation strategies produce a large number of ideal profiles, almost always greater than six, that represent best fitting representations of the data, while at the same time making it impossible to interpret what membership in, say, four or more profiles actually means. This conflict between statistical goodness-of-fit and subject-matter-based interpretability of shared membership cannot usually be resolved using conventional mixed membership models. We show that by introducing separate mixed membership models, each containing a small number of ideal profiles, to describe a population according to responses focused on distinct subject matter domains, and at the same time producing a vector of correlated grade of membership scores for the individuals, interpretation of shared memberships across the distinct subject matter domains becomes feasible. Deciding on what constitutes a good model requires tradeoffs between statistical goodness-of-fit criteria and frequently non-quantifiable subject-matter-based interpretation. We illustrate these unavoidable tradeoffs in several epidemiological contexts.

8.1 Introduction

Mixed membership models are ideally suited for characterizing heterogeneous populations where many individuals have multiple characteristics of interest, no combination of which occur at high frequency in an overall population. Examples of this phenomena are: (i) representation of the health status of elderly populations at the community level (Berkman et al., 1989); (ii) prevalence of, and variation in, co-infection with tropical diseases at the district level (Keiser et al., 2002; Raso et al., 2006); (iii) characterization of environmental and behavioral risks for Chagas disease in Latin America (Chuit et al., 2001); (iv) representation of malaria risk in complex eco-systems (e.g., the Brazilian Amazon) (Castro et al., 2006); and (v) represention of variation in disability among the elderly in the United States (Manton et al., 2006). Although mixed membership models have been fit to data from the above and other settings, very little attention has been given to the severe lack of interpretability of 'shared membership' of characteristics of individuals among a family of extreme, or ideal, types. Indeed, most published analyses to-date using this class of models focus on interpretation of ideal type profiles, and neglect the more complex stories inherent in the phenomenon of shared membership. This is tantamount to using a mixed membership model as a crisp clustering methodology and bypassing the analysis of shared membership, which is the key distinguishing feature of the technology.

Standard estimation strategies for mixed membership models, if used in an exploratory mode, produce a number K of ideal profiles that lead to the best fitting and parsimonious representations of the multidimensional data. The optimum value of K can be defined based on a measure of relative goodness-of-fit of models run for several K values (Manton et al., 1994; Airoldi et al., 2010; White et al., 2012; Suleman, 2013). However, interpreting these profiles can be rather complex. Assuming K ideal profiles, each individual is assigned a grade of membership (GoM) vector **g** that corresponds to coordinates in a simplex with K vertices, each of which represents an ideal, or extreme, type of individual. For example, in the case of two profiles $(K = 2)$, non-zero entries in **g** define the individuals that belong to the k_1 profile, while those equal to zero determine who does not belong to the same k_1 profile, but to k_2 instead. These cases can be understood as vertices (endpoints) of a line. However, some elements do not lie on the vertices but on the line connecting them. They share characteristics of both profiles. If, for example, the vertices correspond to 'very high risk' and 'very low risk', respectively, then $\mathbf{g} = (g_1, 1 - g_1)$ with $0 < g_1 < 1$ identifies an individual with a degree of risk that is intermediate between the extremes corresponding to the vertices. In the case of three profiles $(K = 3)$, however, the geographical representation moves from a line to a triangle, and individuals with shared membership will lie on either an edge of the triangle or in the interior. Individuals on edges share conditions with only two vertices, and degree of similarity to one or the other of them provides for straightforward interpretation. However, individuals in the interior of the triangle share conditions with all three vertices, and writing a coherent English sentence describing the shared conditions becomes a more substantial challenge.

More formally, the number of non-zero entries in **g** is the number of vertices with one or more associated conditions that represent the state of a given individual. For a given K, there are $\binom{K}{2}$ edges of the simplex that represent shared conditions among two vertices. There are $\binom{K}{3}$ faces of the simplex that represent shared conditions among three vertices, and a **g**-vector with all non-zero entries corresponding to an individual who shares conditions with all K vertices. The central problem associated with values of $K \geq 4$ is writing an interpretable description of what it means to share conditions with four or more ideal types. This, of course, is not a statistical problem, but it presents a challenge to investigators that, to the best of our knowledge, has received almost no attention in the extant literature on mixed membership models.

This paper addresses this issue. Our purpose is to present examples of problems where mixed membership modeling with non-standard model specifications and/or vertex-edge-face aggregation schemes facilitates interpretability of shared conditions. Section 8.2 contains specifications of the mixed membership model used in Sections 8.3 and 8.4. In particular, a non-standard specification is introduced where the original response vector is partitioned into blocks of variables associated with distinct subject-matter domains and a mixed membership model is estimated for each domain, thereby generating a set of GoM vectors for each individual, one vector for each domain in the partitioning. This facilitates interpretability in characterizations of disease risk, as we indicate in the next section. In Section 8.3 we discuss an example of characterization of community risk of Chagas disease in rural Argentina, where the original response vector is partitioned into components focused on indices of blood availability and environmental characteristics. Thus, GoM score vectors are generated for each of these domains and interpreted in the context of overall community risk of transmission of Chagas disease. In Section 8.4, we illustrate a standard mixed membership model with 6 pure types, and introduce a vertex-edge aggregation strategy in a study of changes in disability over time in the U.S. population during the period 1982–1994. The aggregation scheme facilitates interpretation of shared membership in a setting where model complexity could, in principle, impair our ability to describe subtleties in the data. We conclude with a brief discussion of open methodological problems that are an outgrowth of our examples.

8.2 Mixed Membership Model Specifications

Let $X = (X_1, \dots, X_J)$ be a vector whose components are variables which can each assume a finite number of possible values. We consider the data analytic task of mapping response vectors $\{X^{(i)}, 1 \leq i \leq N\}$ for N individuals into a unit simplex with K vertices (to be estimated) and a GoM vector, $\mathbf{g}^{(i)} = (g_1, \dots, g_K)^{(i)}$ with $g_1 + \ldots + g_K = 1$ for each individual, specifying a location in the unit simplex. Each vertex of the simplex is associated with a set of levels on a subset of the variables in X. Each set of such levels is interpreted as an ideal, or pure type, set of characteristics. GoM vectors that have all components equal to zero except one of them, say the kth—which must be a 1—identify individuals with response vectors having all of the conditions in the kth pure type. GoM vectors with two or more non-zero entries identify individuals whose response vectors share conditions with the pure types corresponding to the non-zero entries. They exhibit mixed membership across the K pure types, and provide the rationale for the terminology, 'mixed membership models.'

In the conventional version of mixed membership models (Manton et al., 1994; Erosheva et al., 2007), we assume that the variables in the response vector X are independent, conditional on the GoM score vector **g**. More formally, we let $X^{(g)}$ denote the response vector for an individual with GoM score vector **g**. Then we introduce the probability model (Singer, 1989):

$$Pr(X^{(g)} = \ell) = \int_{S_K} Pr(X^{(g)} = \ell | \mathbf{g} = \gamma) d\mu(\gamma) = \int_{S_K} \prod_{j=1}^{J} Pr(X_j^{(g)} = \ell_j | \mathbf{g} = \gamma) d\mu(\gamma), \tag{8.1}$$

where $\mu(\gamma)$ is the distribution of GoM scores and $S_K = \{\gamma = (\gamma_1, \dots, \gamma_K) : \gamma_k \geq 0, \sum \gamma_k = 1\}$ is the unit simplex with K vertices. The conditional probabilities in Equation (8.1) can be written as

$$Pr(X^{(g)} = \ell_j | \mathbf{g} = \gamma) = \sum_{k=1}^{K} \gamma_k \lambda_{k,j,\ell_j},$$

where λ_{k,j,ℓ_j} (called pure type probabilities) can be defined as the probability in profile k of observing level ℓ_j on variable X_j. They are subject to the following constraints (Woodbury and Manton, 1982):

$$\sum_{\ell_j \in L_j} \lambda_{k,j,\ell_j} = 1 \text{ for } 1 \leq k \leq K \text{ and } 1 \leq j \leq J,$$

and L_j is the set of possible levels of variable X_j. Values of these probabilities close to 0 or 1 imply that a distinguished level is either almost certain to appear, or almost never to appear, in the kth pure type. The estimation problem for specification (8.1) is to find K, the associated pure type probabilities, and the individual GoM score vector **g** such that the model provides a good numerical fit to the data and is also interpretable. The words 'good fit' are associated with numerical goodness-of-fit criteria that are context free. However, the word 'interpretable' is not connected to statistical methodology, but is directly linked to our ability to describe the position of an individual in the simplex, i.e., describe the GoM score vector **g** in coherent English sentences that make sense in the scientific setting of a given dataset. The fundamental problem under discussion in this chapter is numerical identification of values of K which can be anywhere from 4 to 60 or 70, depending on the dataset, and the inability of the investigator to write a coherent paragraph describing GoM vectors with four or more non-zero entries and the sharing of membership among many pure types.

Depending on the dataset, and with a large value of K associated with the best fitting model, we may find that nearly all individuals have GoM vectors that place them on a vertex or on an edge of the simplex, sharing conditions with at most 2 pure types. This is a relatively easy situation in which to provide interpretable descriptions of response vectors sharing conditions between the pure types. If 90% or more of the individuals share conditions with at most 3 pure types—i.e., they are, at worst, situated in a face of the simplex—interpretable summaries of the shared conditions can also frequently be produced. In a study of classification of scientific papers (Airoldi et al., 2010), mixed membership models with $K = 20$ were utilized on the basis of goodness-of-fit criteria, but nearly all papers in the classification exercise shared conditions with at most 5 pure types. Most of the papers with shared pure types involved only 2 or 3 such profiles. In the context of that study, even sharing of 3 or 4 pure types resulted in interpretable formulations of the shared conditions. However, GoM vectors with 5 non-zero entries seem to be close to an upper bound on shared condition interpretability. In Section 8.4, we will show an example with $K = 6$, but where a subject-matter-driven aggregation of edges in the unit simplex leads to interpretability of the set of all shared conditions among pairs of vertices.

In studies of disease risk, as illustrated in Section 8.3, the variables in X can frequently be partitioned into subsets, each of which is associated with a different domain of risk. This context also makes it desirable to use 2 pure type specifications corresponding to the extremes of high and low risk for each domain. What we frequently find with high-dimensional X is that 2 pure type models provide a poor fit to the data, but models with 3, 4, and 5 pure types do much better numerically, while paying a high price in losing gradation of risk interpretations of shared conditions. One route out of this dilemma is to change the conditioning structure of the mixed membership model and produce two or more sets of 2 pure type representations, one for each substantively defined risk domain. Then for the variables associated with each domain, we have high and low risk interpretations of 2 pure types and a separate GoM score vector for each individual, one vector for each domain. Essentially we are trading higher dimensionality in K, with a single GoM vector, for lower dimensionality in K (namely, $K = 2$), but with a set of correlated GoM vectors for each individual, one vector for each risk domain.

We describe a mixed membership model structure for a multiple domain specification in the simplest setting (two domains). Let $X = (X^{(1)}, X^{(2)})$ be a response vector with subsets of variables $X^{(i)}$ associated with the subject matter domains indexed by $i = 1, 2$. We introduce the pair of GoM vectors $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ and assume that the variables in each of $X^{(i)}$ for $i = 1, 2$ are independent, conditional on $\mathbf{g} = (\mathbf{g}^{(1)},\ \mathbf{g}^{(2)})$. Then we set

$$\begin{aligned} Pr(X = 1|\mathbf{g} = \gamma) &= \prod_{i\in(1)} Pr(X_i^{(1)} = \ell_i^{(1)}|\mathbf{g} = \gamma) \prod_{i\in(2)} Pr(X_i^{(2)} = \ell_i^{(2)}|\mathbf{g} = \gamma), \\ &= \prod_{i\in(1)} \sum_{k\in K(1)} \gamma_k^{(1)} \lambda_{k,i,\ell_i}^{(1)} \prod_{i\in(2)} \sum_{k\in K(2)} \gamma_k^{(2)} \lambda_{k,i,\ell_i}^{(2)}. \end{aligned} \tag{8.2}$$

Here, $\mathbf{g} = (\mathbf{g}^{(1)},\ \mathbf{g}^{(2)}), K(j) =$ Number of pure types in group$(j), j = 1, 2$. In the context of the risk profiles mentioned above, we would impose $K(1) = K(2) = 2$. In Section 8.3, we present an analysis of Chagas disease risk where the representation (8.2) plays a central role.

8.3 Chagas Disease Risk In Rural Argentina

Rural communities that are endemic for Chagas disease usually consist of privately owned habitats containing a primary house and peridomestic structures that serve as animal pens/housing, crop storage areas, and tool sheds and storage areas for agricultural equipment. The physical characteristics of houses and peridomestic structures are highly variable, thereby creating great heterogeneity within a community in sources of blood meals for triatomine bugs. Dogs, cats, chickens, and young children are all part of the transmission system, and their physical proximity during the night is an important factor in attracting *Triatoma infestans* vectors, the primary transmitters of *Trypanosoma cruzi* parasites that are the causative agents of Chagas disease.

We consider a community with 445 habitats in rural Santiago del Estero, Argentina (Paulone et al., 1991), where at baseline, 99.6% (443/445) habitats were infested with *T. infestans* bugs. As part of the initial data collection, *T. infestans* were collected from 390 (88%) of the houses and from the peridomestic structures of 280 (63%) habitats. A total of 6,518 *T. infestans* were captured in the 390 infested bedroom areas. Of these, 2,249 bugs were examined for *T. cruzi* parasites, and 697 (31%) were found positive. On the human side of the transmission system, 2,153 (69%) of the 3,194 persons in the community were serologically tested. The prevalence rate of seropositivity against *T. cruzi* infection was 29.2% (630/2153). Age specific seropositivity ranged from 9.6% in children under the age of 5 years to 57.7% in persons aged 70 or more. For the age group of children aged 5–14 the seropositivity rate was 25.3%.

Despite the high overall seropositivity rate, there were sets of habitats with very low rates and other sets with high rates in children in the age range 5–14. This is a useful age range for assessing Chagas disease incidence in the relatively recent past, particularly in a community where there has been no active control activity against infestation prior to the study by Paulone et al. (1991). Further, it is not at all obvious which habitats are at highest or lowest risk for Chagas disease transmission on the basis of a walking tour of the community, or even a verbal estimate from the community health officer. Our analytical problem is to identify the highest and lowest risk habitats in the midst of a highly heterogeneous endemic community, with the longer term objective of adapting the features of the low risk habitats on a wider scale as a hopefully low cost means of preventing transmission of *T. cruzi*.

To this end, a mixed membership modeling exercise using specification (8.2) was carried out using variables from two distinct domains: indices of blood availability and characteristics of the physical environment (Chuit et al., 2001). These are delineated in Table 8.1. For each domain a 2-pure type model (interpreted as levels of high and low risk) was fit to the data from the 445 habitats. A GoM score vector associated with each domain was generated for each habitat. Then, with only two profiles for each domain, the GoM vectors $\mathbf{g}^{(1)} = (g_1, 1 - g_1)$ and $\mathbf{g}^{(2)} = g_2, 1g_2)$ associated with each habitat have g_i indicating the degree of similarity of the habitat characteristics to the high risk profile for $i = 1$ (blood availability) and $i = 2$ (environmental characteristics). Using tertile cut points, for each risk domain (blood availability and environmental characteristics) we define a habitat to be low risk if $0 \leq g_i \leq 0.20$; intermediate risk if $0.20 < g_i \leq 0.70$; and high risk if $g_i > 0.70$. Cross classifying the habitats by their scores g_i, for $i = 1, 2$, Table 8.2 shows the breakdown of seropositivity rates for children in the age range 5–14 years.

TABLE 8.1
Response variables used in the mixed membership model.

Indices of Blood Availability	**Conditions**
number of dogs	= 2; > 2
number of cats	= 2: > 2
number of persons	= 2; > 2
persons/room	= 2; > 2
people/bed	= 2; > 2
people/[structures in the habitat]	= 2; > 2
persons + dogs + cats	= 5; 6 - 8; > 8
[persons + dogs + cats]/[room in the house]	= 6; >6
[persons + dogs + cats]/bed	= 4; > 4
[persons + dogs + cats]/[structure in the habitat]	= 3; > 3
Environmental Variables	**Conditions**
seasonal migration	No; Yes
condition of interior roof	Good (cement, zinc, fibrocement); Bad (straw, jarilla, discard)
condition of interior walls	Good (cement, lasterwall, no cracks); Bad (unplastered mud or brick with cracks)
condition of gallery roof	Good; Bad
number of rooms	= 1; > 1
number of beds	= 3; > 3
corn storage area	No; Yes
kitchen	No: Yes
equipment store room	No; Yes
corral	No; Yes
pig pen	No; Yes
brick pile	No; Yes

Source: Chuit et al. (2001).

TABLE 8.2
Seropositivity rates (%) by habitat risk for children aged 5–14.

Risk Domain		**Environmental Characteristics**		
	Risk level	*Low*	*Medium*	*High*
Blood Availability	*Low*	18.8	16.2	21.4
	Medium	20.5	22.0	9.0
	High	19.3	22.8	25.0

Estimation in the GoM model (8.2) was carried out by estimating the GoM score vectors and pure type probabilities for each domain (environmental and blood availability) separately via GoM model (8.1) with $K = 2$ in each case. The two sets of GoM scores, $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$, with a pair of such scores for each habitat, are—not surprisingly–correlated. Determination of the association between GoM scores begins with a scatter plot of (g_1, g_2) for the full set of habitats. Division of GoM scores into tertiles for purposes of classifying habitats by gradations of risk was the result of judgment by the investigators that such coarse graining led to qualitatively different categories of risk in both the environmental and blood availability domains. Quartile, quintile, or finer divisions could have been used, but these do not lead to meaningfully distinct risk categories. We emphasize here that this is not a statistically driven categorization. It is based on subject matter interpretations of meaningfully different levels of risk.

Turning to the profiles per se, the high risk profile for blood availability is represented by the logical AND statement: [more than 2 dogs] AND [6 persons or dogs or cats per room] AND [3 persons or dogs or cats per structure at the habitat]. Low risk is characterized by [None of the adverse conditions in the blood availability section of Table 8.1]. The high risk profile for physical environmental characteristics is given by: [poor interior roof] AND [poor gallery roof] AND [presence of a pig pen] AND [presence of a brick pile]. Low risk for environmental characteristics is characterized by no adverse house or peridomicile conditions from the full list in Table 8.1 (Chuit et al., 2001).

If the pattern of seropositivity rates corresponds to the risk levels for the habitats, Table 8.2 should be a double-gradient table in the sense that the rates should increase in going from left to right across each row and from top to bottom down each column. Rows 1 and 3 and column 2 exhibit this pattern, but there is one exceptional cell (row 2, column 3) corresponding to medium risk on blood availability and high risk on environmental characteristics. This requires some explanation, which we provide below. Column 1 appears to have a violation in row 3; however, these rates (19.3% and 20.5%) are not statistically significantly different from each other at level 0.05.

The aberrant cell (row 2, column 3) is high risk on environmental characteristics. This particularly means that there is a poor interior and gallery roof. While these conditions characterized habitats with high risk environmental characteristics up to approximately 18 months—two years prior to serological data collection for the present study, the owners engaged in roof repair on their houses. The immediate effect was to eliminate localities that were previously hospitable to *T. infestans*. It is, therefore, not surprising that the incidence rate for new Chagas disease cases dropped precipitously at those sites. It is important to emphasize that all owners of houses in the community did not engage in roof repair. Indeed, examination of Table 8.2 provoked a deeper inquiry into why the $(2, 3)$ cell was so anomalous. Examining the full information set for the habitats in this cell revealed that the GoM analysis had isolated the locations in a highly heterogeneous community where roof repair was making a major difference in Chagas disease incidence.

With the extreme habitats identified—meaning those scoring high risk on both blood availability and environmental characteristics as well as those scoring low risk on both dimensions—a more in-depth analysis was carried out to characterize the most (and least) risky habitats. To this end, variables defining host availability and environmental conditions (shown in Table 8.1) were used to calculate the odds ratio (OR) comparing the proportion of habitats with the highest level of risk to the proportion having the lowest level of risk; 95% confidence interval (CI) on the odds ratio was also calculated. A condition was then defined to be extremely risky if the lower bound of the 95% CI on the odds ratio exceeded 3.5. Analogously, for the habitats classified as low risk on both domains in Table 8.2, the odds ratio comparing the lowest risk condition on each variable with the highest risk condition on that variable and its 95% CI was calculated. Now, a condition was defined to be extremely low risk if the lower bound of the 95% CI exceeded 3.5. Applying these stringent criteria, a new set of low and high risk conditions were specified. They are described in Table 8.3 together with the seropositivity rates for the subset of habitats satisfying them.

Identification of habitats with these very different seropositivity rates that were, nevertheless, embedded in an endemic community provided evidence that our mixed membership methodology,

TABLE 8.3
High and low risk conditions and associated seropositivity rates for children aged 5–14.

Level of risk	Conditions	% seropositive
Low	[# of peridomicilliary structures = 1, but no presence of a food storage area] AND [1 dog OR 1 cat]	7.7
High	[# of peridomicilliary structures > 2] AND [presence of food storage area] AND [> 1 dog OR > 1 cat OR both]	36.4

together with the second stage screening of habitats that were low and those that were high on risk variables from the two domains, deserves attention in other risk assessment settings. The low risk conditions in Table 8.3 are associated with a seropositivity rate among children aged 5–14 of 7.7%, significantly lower than the rate observed in the total population (25.3%). These conditions are, in fact, a basis for relatively simple and inexpensive restructuring of individual habitats to substantially reduce Chagas disease risk. Further, from a methodological perspective, the standard mixed membership model structure, which contains only a single GoM score vector, automatically masks over the risk domain distinctions achievable with the partitioned model of Equation system (8.2).

A final methodological point pertaining to this example concerns the bivariate distribution of the GoM scores (g_1, g_2). An empirical distribution is obtained via conditional likelihood calculation of GoM scores for each of the domains separately using the specification (8.1). There is currently no theoretical basis for a priori imposing a class of bivariate distributions to represent the GoM scores from two domains in the context of Chagas disease epidemiology. This situation could change with particular applications, but thus far there is not enough experience using specification 2, or models with three or more distinct domains, to warrant putting forth defensible classes of bivariate distributions for (g_1, g_2). Carrying the modeling into a Bayesian framework would require specification of defensible prior distributions on the GoM scores. Except for a nearly uniform prior on the unit square, we await the development of subject-matter-driven specification of more informative priors for use with model specification (8.2).

8.4 Disability Change In The U. S. Population: 1982–1994

Populations aged 65 and older at the level of communities contain many people who have multiple disabilities and chronic conditions, no combination of which occurs at high frequency. This makes classification of elderly populations into disability/chronic conditions groups particularly problematic. Simply describing the joint distribution of co-morbid conditions is an unwieldy and difficulty task. This setting, however, is precisely where mixed membership models can play a useful role in terms of representing the heterogeneity in elderly populations via interpretable sets of pure types and characterizations of shared membership between them among selected sub-populations. Berkman et al. (1989) put forth an initial analysis in this direction, focused on the elderly community of New Haven, CT.

Data used for the analysis was derived from the National Long Term Care Survey (NLTCS) list-based samples of approximately 20,000 persons age 65+ drawn from Medicare enrollment files in the years 1982, 1984, 1989, and 1994. To ensure a national sample of the age 65+ population at each survey date, a fresh supplementary list sample was drawn from Medicare enrollment files in 1984, 1989, and 1994. A detailed description of the NLTCS is given in Corder et al. (1993). The analysis examined sub-groups that contributed to an overall decline in disability between 1982 and 1994, and some that did not follow this general trend. Mixed membership models with the structure of Equation system (8.1) were fit for each of the survey years to response vectors whose coordinates described the ability, or not, of individuals to perform a diverse set of "Activities of Daily Living" (ADLs), tests of physical functioning, or both. A battery of 27 ADLs, "Instrumental Activities of Daily Living" (IADLs), and functional impairment measures were employed for this purpose. They are listed in Table 8.4.

TABLE 8.4
Activities of daily living and measures of physical functioning assessed in the National Long Term Care Survey.

ADL Items: need help with	IADL Items: need help with
Eating	Heavy work
Getting in/out of bed	Light work
Dressing	Laundry
Bathing	Cooking
Using a toilet	Traveling
Getting about outside	Grocery shopping
	Managing money
Are you	Taking medicine
Bedfast	Telephoning
Using a wheelchair	
Restricted to no inside activity	
Can you	
See well enough to read a newspaper	
How much difficulty do you have: none, some, very difficult, cannot at all	
Climbing 1 flight of stairs	
Bending for socks	
Holding a 10 lb. package	
Reaching over head	
Combing hair	
Washing hair	
Grasping small objects	

Source: Manton et al. (1998).

For the community population, the best fitting mixed membership models for each of the survey years, satisfying Equation system (8.1), had $K = 6$ pure types. There was very little variation in the pure types across the survey years. Independent of the model, a 7th pure type/profile was added for the elderly institutionalized population. This group was quite homogeneous, having an average of 4.8 ADLs chronically impaired. The full set of pure types is described in Table 8.5.

Although the pure types are clearly interpretable, sharing of conditions across sets of 2 and

TABLE 8.5
Disability pure types from mixed membership model with $K = 6$.

I	Active, no functional impairments.
II	Very modest impairment, some difficulty climbing stairs, lifting a 10 lb. package, and bending for socks (no ADL or IADL).
III	Moderate physical impairment, great difficulty climbing stairs, lifting 10 lb. package, reaching over head, etc. (no ADL or IADL).
IV	All IADLs, great difficulty climbing stairs, and lifting a 10 lb. package.
V	Some ADLs and IADLs, difficulty climbing stairs, cannot lift a 10 lb. package.
VI	All ADLs and all IADLs, and all tasks (high percentage in wheelchairs).
VII	Institutionalized – these are not included in the mixed membership model.

Source: Singer and Ryff (2001).

especially 3 pure types with only mild differences among some of the characteristics presents serious difficulties for differentiating among sub-groups. To resolve this difficulty, we introduce a context-specific strategy for aggregating vertices and edges of the simplex to create a coarser set of disability categories. For this, observe that pure types I–III represent persons who are generally functionally intact. In contrast, pure types IV–VI identify persons with significant physical or cognitive impairments. Heterogeneity within the functionally intact group is represented by persons who share conditions with pairs of pure types I–III. Such people have GoM score vectors at a given survey with non-zero entries for precisely 2 pure types. For example, $\mathbf{g} = (0.3, 0.7, 0, 0, 0, 0)$ is the GoM score vector for a person whose responses on ADL, IADL, and physical functioning are closer to pure type II (a weighting of 0.7) than to pure type I (a weighting of 0.3). We will denote the category of functionally intact persons by C(1 - 3). There are persons with response vectors at one of the pure types I, II, or III, supplemented by persons who share conditions with any pair of them. Geometrically, persons in C(1 - 3) are either at one of the vertices in the unit simplex labeled I, II, or III, or they are on one of the edges that link pairs of these vertices.

Heterogeneity in the severely disabled group, labeled C(4 - 6), is represented by persons at pure types IV, V, and VI, or by those who share conditions with any pair of them. A different form of heterogeneity, C(int), is designated for persons who are on edges connecting one of the vertices [I, II, III] to one of the vertices [IV, V, VI]. A more extreme form of heterogeneity, designated C(res) is represented by persons who share conditions with 3 or more pure types. Geometrically they are identified by points in the faces or further in the interior of the unit simplex with $K = 6$. The partitioning of population aged 65+ into the four disability categories defined above, augmented by the institutionalized population, identifies clearly distinct groups with qualitatively different interpretations of their mix of disabilities.

Returning to the issue of disability decline mentioned at the beginning of this section, we know that there was a decline of 1.5% per annum in the proportion of the age 65+ population that was chronically disabled over the time period 1982–1994. The classification scheme introduced above facilitates our getting a much better picture of the variation in prevalence of chronic disability according to our more refined classification of it. To this end, Table 8.6 shows the percent per annum changes in prevalence of chronic disability from 1982–1994 by disability category generated from the aggregation of vertices and edges of the unit simplex that gave rise to C(1 - 3), C(4 - 6), C(int), and C(res). The table also includes the category Inst, which refers to institutionalized individuals.

Prior to the production of Table 8.6, separate GoM models were run for the years 1982, 1984,

TABLE 8.6
Percent per annum changes in prevalence of chronic disabilities from 1982–1994 by age and gender.

Disability	Men aged	Women aged	Men aged	Women aged
category	65 - 84	65 - 84	85+	85+
C(1 - 3)	+0.21	+0.30	+1.45	+0.25
C(4 - 6)	-2.78	-5.31	-2.65	-2.91
C(int)	-2.11	-0.74	+4.08	-0.16
C(res)	-1.05	-1.01	-3.13	+0.05
Inst.	-1.60	-1.71	-0.94	+0.16

Source: Singer and Ryff (2001).

1989, and 1994. The remarkable feature of these separate analyses was that the number of pure types and the conditions entering into them were invariant over this 12-year period. Thus, changes involving an individual were captured in changes in their GoM score vectors over time. Having a chronic disability means having at least one ADL or IADL, where such disability has lasted, or was expected to last, at least 90 days. Prevalence of this condition in each of the disability categories is the basis for calculation of changes between 1982 and 1994.

It is important to emphasize that the invariance in number of pure types and the conditions that enter into them that were an empirical fact of life in the present analysis is by no means generic to GoM modeling over time. Indeed the number of pure types could have varied between, for example, 3 and 7, depending upon the time of assessment. In addition, the conditions entering into pure types at each assessment time could be different. Under such a scenario, it would be impossible to discuss changes over time via a table as simple as Table 8.6.

In-depth interpretation of the category-specific changes in Table 8.6 requires the use of a much richer set of variables from the NLTCS then used for the present methodological discussion of mixed membership models. Extensive analysis of disability changes can be found in Manton et al. (1998), Manton et al. (2006), and Manton (2008).

8.5 Discussion And Open Problems

The major feature of mixed membership models that motivated their specification in the first place (Woodbury and Clive, 1974; Woodbury et al., 1978) was the empirical fact, arising in many studies, that crisp classification of individuals into well-defined categories was frequently difficult, if not impossible. Standard clustering methods do not provide a way out of this impasse, and the observation that shared membership among two or more categories for individuals in a wide variety of scientific contexts is conceptually meaningful paved the way for elaboration of formal models to capture this idea (Woodbury and Clive, 1974; Woodbury et al., 1978; Davidson et al., 1989). Although mixed membership models can be specified according to a priori theories and used in a hypothesis testing mode, by far the most extensive use of the methodology has been in exploratory studies where K, the number of pure types, and the structure of the pure types themselves, is estimated from the data. In terms of numerical goodness-of-fit criteria, best fitting mixed membership models have been obtained in many instances where K takes on values in the range 15–50 (Airoldi et al., 2010). Then interpretive reports are presented with a focus on the structure of the pure types themselves, with only minimal—if any—discussion and interpretation of shared membership across 2 or more pure types. For a notable exception to this practice, see Airoldi et al. (2010).

Our own attempts to consider shared membership (Berkman et al., 1989; Chuit et al., 2001; Castro et al., 2006) rather quickly highlighted the interpretability difficulties involved in simply writing coherent sentences to explain shared membership involving 4 or 5 pure types. This led to the alternative specification shown in Equation (8.2), which is the simplest example of partitioning response vectors by distinct subject-matter domains and the introduction of the assumption of conditional independence of variables in each domain separately given the set of GoM score vectors for all of them. The problem of interpreting shared conditions across multiple pure types is then transferred to one of providing interpretable explanations for the correlation structure in the set of GoM score vectors across domains. The latter situation turned out to be especially informative in our example of characterizing Chagas disease risk in rural Argentina. This special setting is also a generic one for disease risk assessment. In fact, we think it highly desirable to apply the strategy implied by Equation (8.2) to represent risk profiles in complex eco-epidemiological contexts quite generally. In particular, the process of health impact assessment (HIA) for large scale industrial projects in the tropics could benefit from this approach (Krieger et al., 2012; Winkler et al., 2011; 2012a;b).

The example of classification of disability in the U.S. population, discussed in Section 8.4, is a prototype for interpreting shared conditions where a multiplicity of pure types can meaningfully be aggregated into coarse categories of roughly similar conditions. It is unclear, in terms of scientific subject-matter, how to characterize the problems that lend themselves to this kind of aggregation methodology. However, we feel it would be useful to attempt such pure type, edge, and even face consolidations in exploratory data analyses using the mixed membership specification shown in Equation (8.1), when K is in the range 4–6, and certainly when $K > 10$.

In summary, we demonstrate alternative specifications of mixed membership models where an increase in dimensionality of grade of membership scores is traded for simplicity in the number and structure of ideal types, with clear payoff in terms of interpretability of model output. Alternatively, sub-sets of vertices and edges—or even faces—linking them can be aggregated to form interpretable categories of individuals, about which coherent descriptions can be formulated. The scientific subject matter must dictate which among these and other dimension-reducing strategies are to be employed for a particular problem. Although the statistical details of fitting mixed membership models to data lie outside the scope of the present paper, we direct the reader to the interesting Bayesian formulations in Airoldi et al. (2008; 2010) that focus on model specification (8.1). There is no analogous rigorous Bayesian methodology to-date for the class of specifications exemplified by Equation (8.2). Here is an important challenge worth taking up in the immediate future.

Acknowledgments

BHS thanks the Army Research Office Multidisciplinary Research Initiative for financial support under MURI Grant #58153-MA-MUR, Prime Award #W91 INF-11-1-0036. MCC thanks the National Institute of Allergy and Infectious Diseases, Award Number R03AI094401-01 (PI Castro). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

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.

Airoldi, E. M., Erosheva, E. A., Fienberg, S. E., Joutard, C., Love, T., and Shringarpure, S. (2010). Reconceptualizing the classification of PNAS articles. *Proceedings of the National Academy of Sciences* 107: 20899–20904.

Berkman, L., Singer, B. H., and Manton, K. G. (1989). Black/white differences in health status and mortality among the elderly. *Demography* 26: 661–678.

Castro, M. C., Monte-Mór, R. L., Sawyer, D. O., and Singer, B. H. (2006). Malaria risk on the Amazon frontier. *Proceedings of the National Academy of Sciences* 103: 2452–2457.

Chuit, R., Gurtler, R. E., Mac Dougall, L., Segura, E. L., and Singer, B. H. (2001). Chagas disease – risk asssessment by an environmental approach in Northern Argentina. *Revista de Patologia Tropical* 30: 193–207.

Corder, L. S., Woodbury, M. A., and Manton, K. G. (1993). Health loss due to unobserved morbidity: A design based approach to minimize nonsampling error in active life expectation estimates. In *Proceedings of the 6th International Workshop on Calculation of Health Expectancies (INSERM)*. Paris, France: J. Libbey Eurotext, 217–232.

Davidson, J. R., Woodbury, M. A., Zisook, S., and Giller, Jr., E. L. (1989). Classification of depression by Grade of Membership: A confirmation study. *Psychological Medicine* 19: 987–998.

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

Keiser, J., N'Goran, E., Traore, M., Lohourignon, K. L., and Singer, B. H. (2002). Polyparasitism in *Schistosoma mansoni*, geohelminths, and intestinal protozoa in rural Côte d'Ivoire. *Journal of Parasitology* 88: 461–466.

Krieger, G. R., Bouchard, M. A., de Sa, I. M., Paris, I., Balge, Z., Williams, D., Singer, B. H., Winkler, M. S., and Utzinger, J. (2012). Enhancing impact: Visualization of an integrated impact assessment strategy. *Geo-Spatial Health* 6: 303–306, + video.

Manton, K. G. (2008). Recent declines in chronic disability in the elderly U.S. population: Risk factors and future dynamics. *Annual Review of Public Health* 29: 91–113.

Manton, K. G., Gu, X., and Lamb, V. L. (2006). Change in chronic disability from 1982 to 2004/2005 as measured by long-term changes in function and health in the U.S. elderly population. *Proceedings of the National Academy of Sciences* 103: 18374–18379.

Manton, K. G., Stallard, E., and Corder, L. S. (1998). The dynamics of dimensions of age-related disability from 1982 to 1994 in the U.S. elderly population. *Journal of Gerontology: Biological Sciences* 53A: B59–B70.

Manton, K. G., Woodbury, M. A., and Tolley, H. D. (1994). *Statistical Applications Usings Fuzzy Sets*. New York, NY: John Wiley & Sons.

Paulone, I., Chuit, R., Pérez, A. C., Canale, D., and Segura, E. L. (1991). The status of transmission of *Trypanasoma cruzi* in an endemic area of Argentina prior to control attempts, 1985. *Annals of Tropical Medicine and Parasitology* 85: 489–497.

Raso, G., Vounatsou, P., Singer, B. H., N'Goran, E., Tanner, M., and Utzinger, J. (2006). An integrated approach for risk profiling and spatial prediction of *Schistosoma mansoni*—hookworm coinfection. *Proceedings of the National Academy of Sciences* 103: 6934–6939.

Singer, B. H. and Ryff, C. D. (2001). Person-centered methods for understanding aging: The integration of numbers and narratives. In Binstock, R. and George, L. K. (eds), *Handbook of Aging and the Social Sciences*. San Diego, CA: Academic Press, 44–65.

Singer, B. H. (1989). Grade of Membership representations: Concepts and problems. In Karlin, S., Anderson, T. W., Athreya, K. B., and Iglehart, D. L. (eds), *Probability, Statistics, and Mathematics: Papers in Honor of Samuel Karlin*. Boston, MA: Academic Press, 317–334.

Suleman, A. (2013). An empirical comparison between Grade of Membership and principal component analysis. *Iranian Journal of Fuzzy Systems* 10: 57–72.

White, A., Chan, J., Hayes, C., and Murphy, T. B. (2012). Mixed membership models for exploring user roles in online fora. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM 2012)*. Palo Alto, CA, USA: AAAI.

Winkler, M. S., Divall, M. J., Krieger, G. R., Balge, M. Z., Singer, B. H., and Utzinger, J. (2011). Assessing health impacts in complex eco-epidemiological settings in the humid tropics: The centrality of scoping. *Environmental Impact Assessment Review* 31: 310–319.

Winkler, M. S., Divall, M. J., Krieger, G. R., Schmidlin, S., Magassouba, M. L., Knoblauch, A. M., Singer, B. H., and Utzinger, J. (2012a). Assessing health impacts in complex eco-epidemiological settings in the humid tropics: Modular baseline health surveys. *Environmental Impact Assessment Review* 33: 15–22.

Winkler, M. S., Krieger, G. R., Divall, M. J., Singer, B. H., and Utzinger, J. (2012b). Health impact assessment of industrial development projects: A spatio-temporal visualization. *Geo-Spatial Health* 6: 299–301, + video.

Woodbury, M. A. and Clive, J. (1974). Clinical pure types as a fuzzy partition. *Journal of Cybernetics* 4: 111–121.

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

Woodbury, M. A. and Manton, K. G. (1982). A new procedure for analysis of medical classification. *Methods of Information in Medicine* 21: 210–220.