Applying Bayesian Mixed Membership Models for Soft Clustering and Classification to Longitudinal Data

Elena Erosheva

University of Washington

and

Stephen E. Fienberg

Carnegie Mellon University

Outline

- Some soft classification problems.
- Mixed membership models.
- Principal Applications:
    - NLTCS disability survey data
      Erosheva (2002); Erosheva-Fienberg (2005)
    - PNAS text and references data
        - Erosheva-Fienberg-Lafferty (2004)
- Generalizations and extensions to account for longitudinal structures.

Ex. 1: NLTCS Disability Data

- **National Long Term Care Survey assesses disability in U.S. elderly population.**
- **2^{16} contingency table with data on functional disability from 1982, 1984, 1989, 1994 waves.**
  - **6 ADLs and 10 IADLs:**

    eating, getting in/out of bed, getting around inside, dressing, bathing, using a toilet, doing heavy house work, doing light house work, doing laundry, cooking, grocery shopping, getting about outside, traveling, managing money, taking medicine, telephoning

Ex. 2: Peanut Butter Brand Choice

- Nielsen scanner panel data.
- 488 households over 4715 choice occasions (at least 5 per HH) for 8 top brands of peanut butter.
- For each choice occasion we have:
  - Shelf price.
  - Information on display/feature promotion.
- Household characteristics used to define "market segments."

Seetharaman, Feinberg, and Chintagunta (2002)
Varki and Chitgunta (2003)
Cooil and Varki (2003)

Ex 3:Matching Words & Pictures

- **Modeling multi-modal data sets, focusing on segmented images with associated text.**

Blei and Jordan (*SIGIR*, 2003)

Barnard, et al. (*J. Machine Learning Research*, 2003)

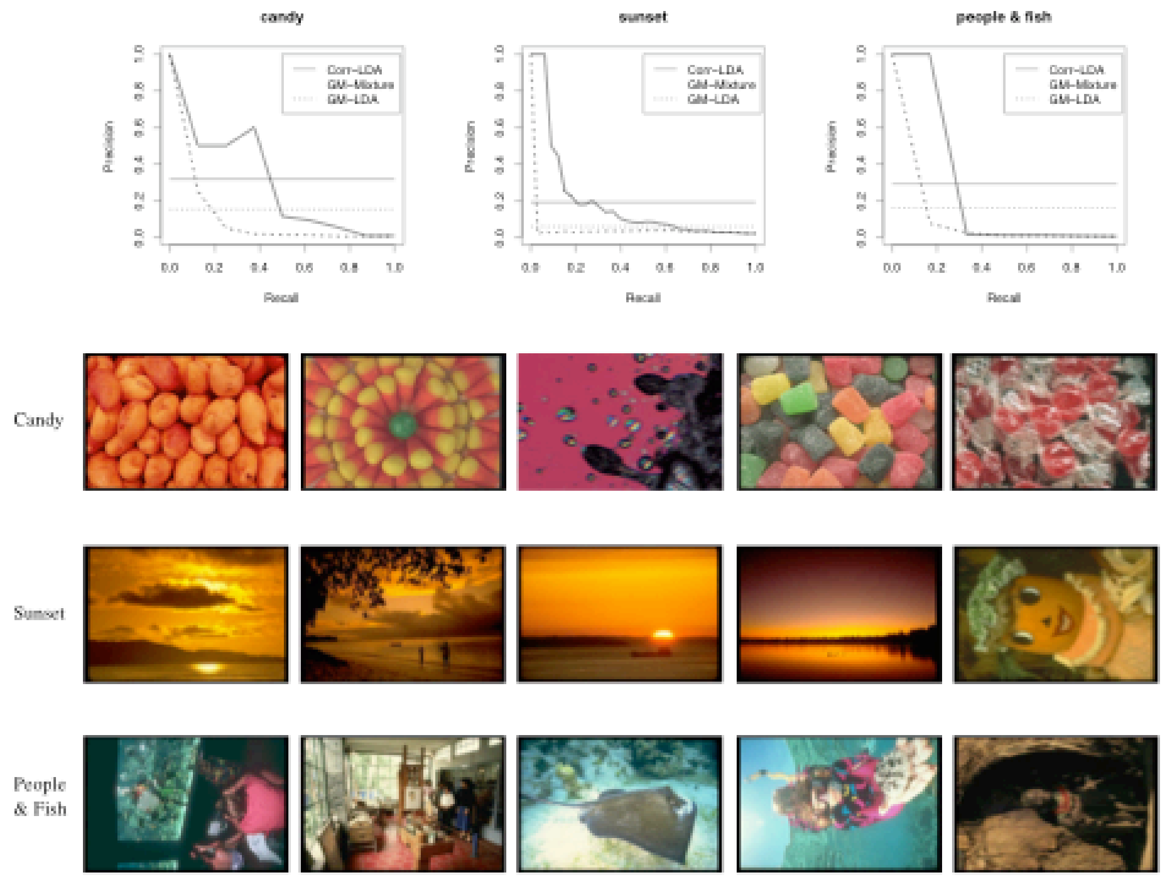


Figure 8: Three examples of text-based image retrieval. (Top) Precision/recall curves for three queries on a 200-factor Corr-LDA model. The horizontal lines are the mean precision for each model. (Bottom) The top five returned images for the same three queries.

Ex. 4: Population Genetic Structure

- **Data on human population structure using genotypes at 377 autosomal microsatellite loci in 1056 individuals from 52 populations.**

Rosenberg, Pritchard, et al. (2002, *Science*)

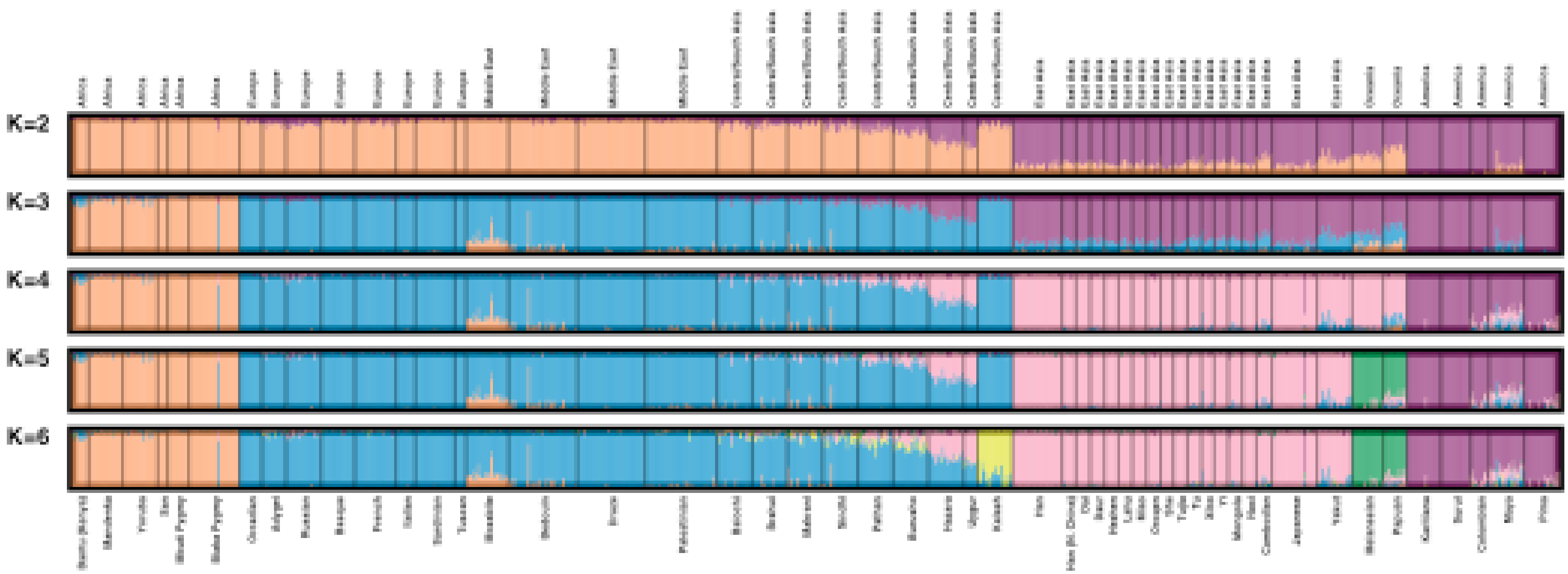


Fig. 1. Estimated population structure. Each individual is represented by a thin vertical line, which is partitioned into K colored segments that represent the individual's estimated membership fractions in K clusters. Black lines separate individuals of different populations. Populations are labeled below the figure, with their regional affiliations above it. Ten *structure* runs at each K produced nearly identical individual membership coefficients, having pairwise similarity coefficients above 0.97, with the exceptions of comparisons involving four runs at $K = 3$ that separated East Asia instead of Eurasia, and one run at $K = 6$ that separated Karitiana instead of Kalash. The figure shown for a given K is based on the highest probability run at that K.

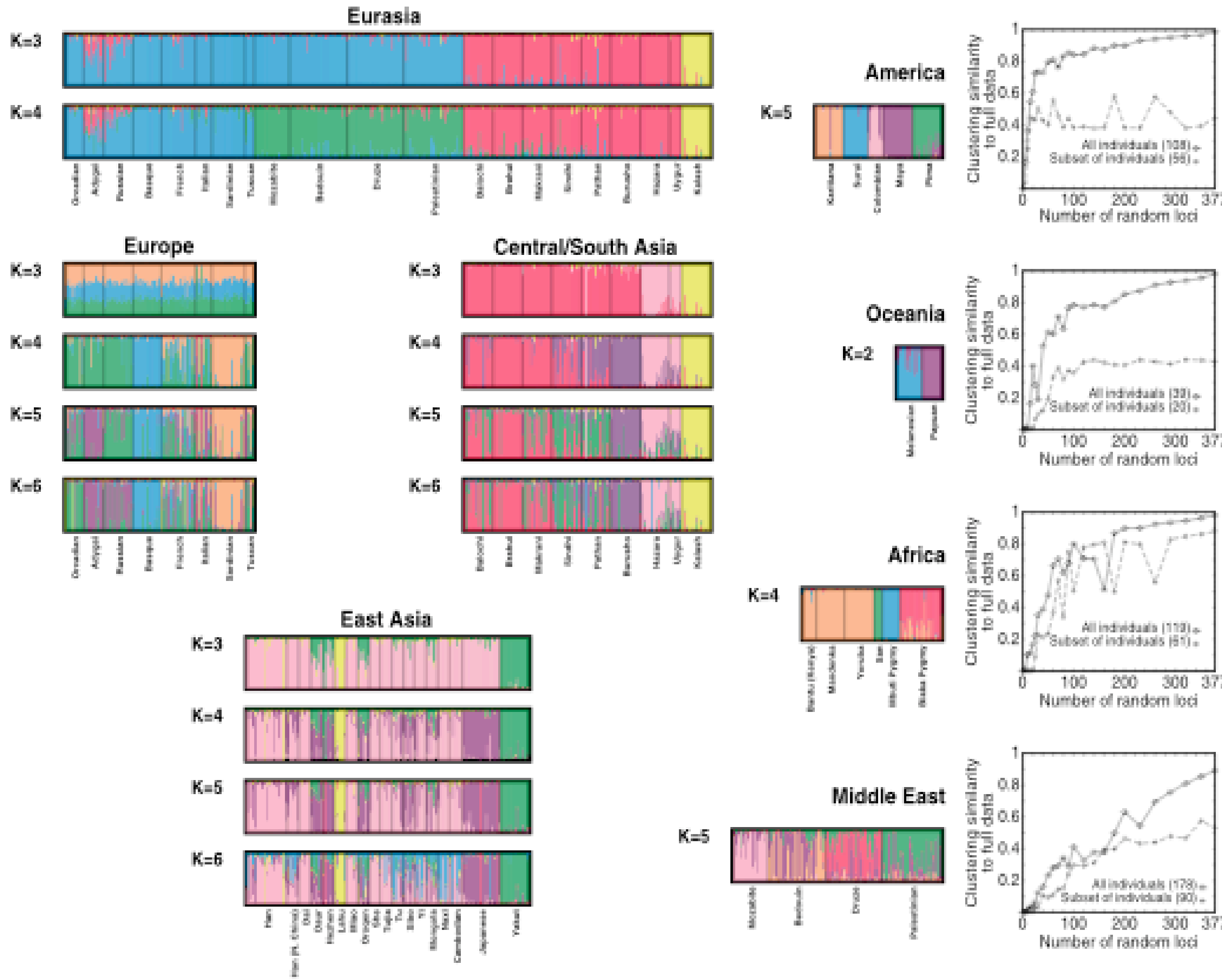


Fig. 2. Estimated population structure for regions. For America, Oceania, Africa, and the Middle East, solutions were consistent across 10 runs (all similarity coefficients above 0.97, 0.93, 0.97, and 0.86, respectively, except those involving one run with Africa that assigned many Biaka individuals partial membership with San). Values of *K* shown for these samples are the highest values for which this was true, and the highest probability runs are shown. For remaining regions, solutions were more variable across runs, and the highest probability runs for various values of *K* are displayed. Graphs for America, Oceania, Africa, and the Middle East display median similarity coefficients between runs based on the full data and runs based on subsets of the data. Correspondence of colors across figures for different regions is not meaningful.

Ex. 5: PNAS Articles

- Proceedings of the National Academy of Sciences U S A.
- **Biological Sciences articles: 92.53% research publications.**
  - **19 subtopics for biological science classification.**
- **Volumes 94-98 (1997-2001):**
  - **39,616 unique words in abstracts.**
  - **77,115 unique references in bibliographies.**

Erosheva, Fienberg, Lafferty (*PNAS*, 2004)
Griffiths and Styvers (*PNAS*, 2004)

Longitudinal Nature of Examples

- NLTCS Disability Data
  - 6 waves of panel data
- Brand choice scanner data
- Genetics
  - looking at multigenerational information
- PNAS articles

Mixed Membership Models

- **Traditional mixture models assume each object belongs exclusively to one of *K* groups or latent classes.**
- **When attributes have mixed origins from different groups, e.g.,**
  - **individual responses in attitude survey,**
  - **words in a scientific article,**
  - **racial origins of people,**

  we have mixed membership.

Mixed Membership Models

- **Hierarchical Bayesian model "membership" represented in terms of weighted combinations of subpopulations ["pure types" or "aspects"].**
- **Assumptions at 4 levels:**
  - **Population level.**
  - **Unit level.**
  - **Latent variable level.**
  - **Sampling scheme.**

Assumption 1: Population Level

- **Population contains *K* subpopulations, with *J* distinct characteristics observed on replicates:**
  - **Observed response pattern $\mathrm{x}_1,\ldots,\mathrm{x}_J$ for subpopulation *k* is characterized by distribution $f(\mathrm{x}_j \mid \theta_{kj})$.**
  - **Response patterns $\mathrm{x}_1^{(r)},\ldots,\mathrm{x}_J^{(r)}$ are independent within each subpopulation.**

Assumption 2: Unit Level

- **We characterize population units by their membership scores:** $\lambda = (\lambda_1, \ldots, \lambda_K)$;
  - **Given membership scores, responses** $\mathrm{x}_1, \ldots, \mathrm{x}_J$ **are independent;**
- **Unit's conditional probabilities are convex combination of corresponding probabilities for *K* subpopulations:**

$$\Pr(\mathrm{x}_j \mid \lambda) = \sum_k \lambda_k \cdot f(\mathrm{x}_j \mid \theta_k).$$

Unit and Population Levels

- **Combination of first two levels or assumptions is equivalent to two-stage process:**
  - ***First stage:* Draw latent classification variable z_j:** $\Pr(z_j = k \mid \lambda) = \lambda_k.$
  - ***Second stage:* Determine distribution of x_{ij} given value of latent classification variable, z_j:** $\Pr(x_j \mid z_j = k) = f(x_j \mid \theta_k).$
- **Averaging over distribution of z_j yields:**

$$\Pr(x_j \mid \lambda) = \sum_k \lambda_k \cdot f(x_j \mid \theta_k).$$

Assumption 3: Latent Variable

- *Random-effects* approach: membership scores are random, i.e., $\lambda \sim D_\alpha$.

$$\Pr(\mathrm{x}_j \mid \alpha;\theta) = \int \left(\sum_k \lambda_k \cdot f(\mathrm{x}_j \mid \theta_{kj}) \right) dD_\alpha(\lambda)$$

Assumption 4: Sampling Scheme

- **Observations on N independent units:**
  - J = # of observed distinct characteristics,
  - R_j = # of replications for jth characteristic.

•If membership scores are independent and drawn at random

$$\Pr\left(\left\{ \mathbf{x}_1^{(r)},\ldots,\mathbf{x}_J^{(r)} \right\}_{r=1}^{R_j} \middle| \lambda,\theta\right)$$

$$= \int \left(\prod_j \prod_r \sum_k \lambda_k f(\mathbf{x}_j^{(r)} \mid \theta_{kj}) \right) dD_\alpha(\lambda).$$

Model for "Survey" Data

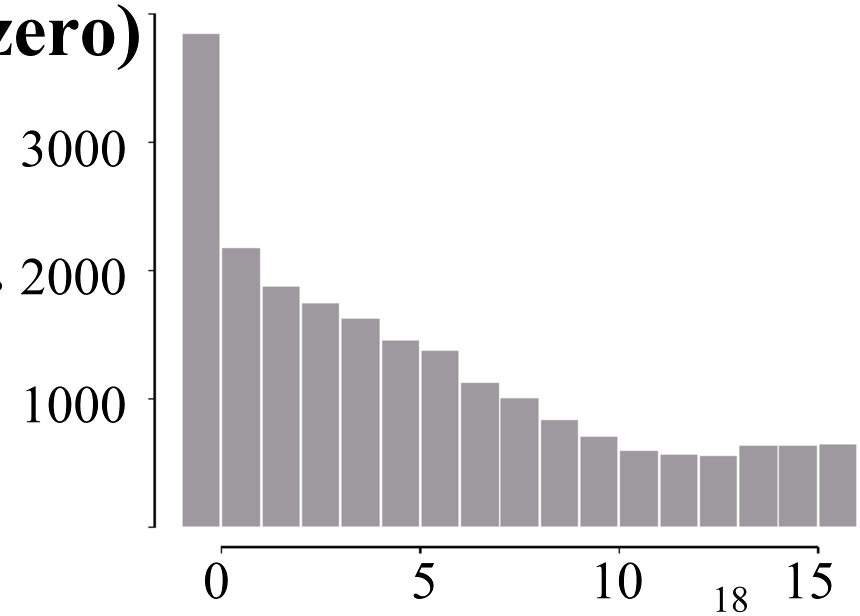
- **Grade of Membership (GoM) model:**
  - **Membership scores define how close individual is to each subpopulations (extreme profiles);**
  - **J dichotomous items, no replications (R=1).**
  - **Probability distribution of jth response, given full membership in kth extreme profile, is**

$$f(x_{ij} \mid \lambda_{ik} = 1; \theta_{kj}) = Binomial(\theta_{kj}) .$$

- **Model has interesting geometric representation.**
  - **Erosheva (2005).**

NLTCS Disability Data

- 2^{16} contingency table with functional disability data from 1982, 1984, 1989, 1994.
  - 6 ADLs and 10 IADLs:
  - J=16; R=1; N=21,574.
  - 65,536 cells (3,152 non-zero)
    - 82% of cell counts < 5.
    - 4% of cell counts > 20.
    - 18% with no disabilities.
    - 3% with all 16.



GOM Implementation

- **Full MCMC to get posterior distribution using Metropolis-Hastings within Gibbs.**
  - **Fit GoM model with *K*=2,3,4, 5 and 9 profiles.**
- **Parameters of interest:**
  - **Conditional response probabilities λ,**
  - **Dirichlet hyperparameters α_0 and ξ.**

GoM Results for 24 Large Cells

n	response pattern	observed	$K=2$	$K=3$	$K=4$	$K=5$	$K=9$
1	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	3853	1249	2569	2055	2801	3651
2	0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0	216	212	225	172	177	218
3	0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0	1107	1176	1135	710	912	1103
4	0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0	188	205	116	76	113	230
5	0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0	122	259	64	88	58	109
6	0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0	351	562	344	245	250	300
7	0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0	206	69	20	23	116	102
8	0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0	303	535	200	126	324	255
9	0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0	182	70	44	71	170	157
10	0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0	108	99	51	39	162	101
11	0 0 1 0 1 0 1 0 0 0 0 1 0 0 0 0	106	16	32	94	94	128
12	0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0	195	386	219	101	160	22
13	0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0	198	369	127	111	108	193
14	0 0 0 0 0 0 1 0 0 0 1 0 1 0 0 0	196	86	41	172	90	187
15	0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0	123	174	96	86	132	98
16	0 0 0 0 0 0 1 0 0 0 1 1 1 0 0 0	176	44	136	162	97	158
17	0 0 1 0 0 0 1 0 0 0 1 1 1 0 0 0	120	9	144	104	41	77
18	0 0 0 0 1 0 1 0 0 0 1 1 1 0 0 0	101	12	127	90	54	72
19	0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0	102	57	44	38	22	95
20	1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0	107	35	88	104	96	37
21	0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0	104	122	269	239	202	60
22	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0	164	55	214	246	272	191
23	0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	153	80	291	261	266	225
24	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	660	36	233	270	362	623
Sum		9141	5917	6829	5683	7079	8392

NLTCS: Choosing *K*

- DIC for GoM model:

K	DIC
2	266912
3	243524
4	229296
5	225323
6	221760

K	DIC
7	215548
8	212108
9	210847
10	210148

- AIC, BIC, and DIC all continue to decrease for latent class model for K= 6,7,8.
- LCM estimation problems occur for K=8.
- We are in process of implementing an approximation to BIC for GoM model.

Model for Scientific Publications

- **Mixed membership for words and references:**
  - **Membership scores are proportions of document's context originating from each aspect.**
  - **J=2 characteristics (words and reference).**
  - **R_i replications vary from document to document.**

PNAS Topical Distribution

	Topic	Number		Percent
1	Biochemistry	2578	(33)	21.517
2	Medical Sciences	1547	(13)	12.912
3	Neurobiology	1343	(9)	11.209
4	Cell Biology	1231	(10)	10.275
5	Genetics	980	(14)	8.180
6	Immunology	865	(9)	7.220
7	Biophysics	636	(40)	5.308
8	Evolution	510	(12)	4.257
9	Microbiology	498	(11)	4.157
10	Plant Biology	488	(4)	4.073
11	Developmental Biology	366	(2)	3.055
12	Physiology	340	(1)	2.838
13	Pharmacology	188	(2)	1.569
14	Ecology	133	(5)	1.110
15	Applied Biological Sciences	94	(6)	0.785
16	Psychology	88	(1)	0.734
17	Agricultural Sciences	43	(2)	0.359
18	Population Biology	43	(5)	0.359
19	Anthropology	10	(0)	0.083
	Total	11981	(179)	100

Years 1997-2001

Unique words: 39,615

Unique references: 77,115

Example 1: PNAS 98(19), 10757-10762.

Heading:

Reward and punishment

Karl Sigmund*†, Christoph Hauert*, and Martin A. Nowak‡§

*Institute for Mathematics, University of Vienna, Strudlhofgasse 4, A-1090 Vienna, Austria; †IIASA, A-2361 Laxenburg, Austria; and ‡Institute for Advanced Study, Einstein Drive, Princeton, NJ 08540

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved May 25, 2001 (received for review March 30, 2001)

Abstract and references:

Minigames capturing the essence of Public Goods experiments show that even in the absence of rationality assumptions, both punishment and reward will fail to bring about prosocial behavior. This result holds in particular for the well-known Ultimatum Game, which emerges as a special case. But reputation can induce fairness and cooperation in populations adapting through learning or imitation. Indeed, the inclusion of reputation effects in the corresponding dynamical models leads to the evolution of economically productive behavior, with agents contributing to the public good and either punishing those who do not or rewarding those who do. Reward and punishment correspond to two types of bifurcation with intriguing complementarity. The analysis suggests that reputation is essential for fostering social behavior among selfish agents, and that it is considerably more effective with punishment than with reward.

1. Gintis, H. (2000) *J. Theor. Biol.* **206,** 169–179.
2. Wedekind, C. & Milinski, M. (2000) *Science* **288,** 850–852.
3. Fehr, E. & Gächter, S. (1998) *Euro. Econ. Rev.* **42,** 845–859.
4. Kagel, J. H. & Roth, A. E., eds. (1995) *The Handbook of Experimental Economics* (Princeton Univ. Press, Princeton, NJ).
5. Bolton, G. E. & Zwick, R. (1995) *Games Econ. Behav.* **10,** 95–121.
6. Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. & McElreath, R. (2001) *Am. Econ. Rev.* **91,** 73–78.
7. Gintis, H. (2000) *Game Theor. Evol.* (Princeton Univ. Press, Princeton, NJ).
8. Henrich, J. & Boyd, R. (2001) *J. Theor. Biol.* **208,** 79–89.
9. Boyd, R. & Richerson, P. J. (1992) *Ethol. Sociobiol.* **13,** 171–195.
10. Clutton-Brock, T. H. & Parker, G. A. (1995) *Nature (London)* **373,** 209–216.
11. Gale, J., Binmore, K. & Samuelson, L. (1995) *Games Econ. Behav.* **8,** 56–90.
12. Huck, S. & Oechssler, J. (1996) *Games Econ. Behav.* **28,** 13–24.
13. Gaunersdorfer, A., Hofbauer, J. & Sigmund, K. (1991) *Theor. Popul. Biol.* **29,** 345–357.
14. Hofbauer, J. & Sigmund, K. (1998) *Evolutionary Games and Population Dynamic* (Cambridge Univ. Press, Cambridge, U.K.).
15. Cressman, R., Gaunersdorfer, A. & Wen, J. F. (2000) *Int. Game Theor. Rev.* **2,** 67–82.
16. Schlag, K. (1998) *J. Econ. Theor.* **78,** 130–156.
17. Weibull, J. W. (1996) *Evolutionary Game Theory* (MIT Press, Cambridge, MA).
18. Nowak, M. A., Page, K. M. & Sigmund, K. (2000) *Science* **289,** 1773–1775.
19. Berger, U. (2001) *Vienna University of Economics*, preprint.

Example 2: PNAS 98(20), 11503-11508.

Heading:

Targeted adenovirus-induced expression of IL-10 decreases thymic apoptosis and improves survival in murine sepsis

Caroline Oberholzer*, Andreas Oberholzer*, Frances R. Bahjat*, Rebecca M. Minter*, Cynthia L. Tannahill*, Amer Abouhamze*, Drake LaFace†, Beth Hutchins†, Michael J. Clare-Salzler‡, and Lyle L. Moldawer*§

Departments of *Surgery and ‡Pathology, Immunology, and Laboratory Medicine, University of Florida College of Medicine, Gainesville, FL 32610; and †Canji Inc., San Diego, CA 92121

Communicated by Charles A. Dinarello, University of Colorado Health Sciences Center, Denver, CO, July 3, 2001 (received for review November 19, 2000)

Abstract and some references:

Sepsis remains a significant clinical conundrum, and recent clinical trials with anticytokine therapies have produced disappointing results. Animal studies have suggested that increased lymphocyte apoptosis may contribute to sepsis-induced mortality. We report here that inhibition of thymocyte apoptosis by targeted adenovirus-induced thymic expression of human IL-10 reduced blood bacteremia and prevented mortality in sepsis. In contrast, systemic administration of an adenovirus expressing IL-10 was without any protective effect. Improvements in survival were associated with increases in Bcl-2 expression and reductions in caspase-3 activity and thymocyte apoptosis. These studies demonstrate that thymic apoptosis plays a critical role in the pathogenesis of sepsis and identifies a gene therapy approach for its therapeutic intervention.

1. Abraham, E., Anzueto, A., Gutierrez, G., Tessler, S., San Pedro, G., Wunderink, R., Dal Nogare, A., Nasraway, S., Berman, S., Cooney, R., *et al.* (1998) *Lancet* **351,** 929–933.
2. Fisher, C. J., Jr., Dhainaut, J. F., Opal, S. M., Pribble, J. P., Balk, R. A., Slotman, G. J., Iberti, T. J., Rackow, E. C., Shapiro, M. J., Greenman, R. L., *et al.* (1994) *J. Am. Med. Assoc.* **271,** 1836–1843.
3. Oberholzer, C., Oberholzer, A., Clare-Salzler, M. & Moldawer, L. L. (2001) *FASEB J.* **15,** 879–892.
4. Hotchkiss, R. S., Swanson, P. E., Freeman, B. D., Tinsley, K. W., Cobb, J. P., Matuschak, G. M., Buchman, T. G. & Karl, I. E. (1999) *Crit. Care Med.* **27,** 1230–1251.
5. Fukuzuka, K., Edwards, C. K., 3rd, Clare-Salzler, M., Copeland, E. M., 3rd, Moldawer, L. L. & Mozingo, D. W. (2000) *Am. J. Physiol.* **278,** R1005–R1018.
6. Zheng, L., Fisher, G., Miller, R. E., Peschon, J., Lynch, D. H. & Lenardo, M. J. (1995) *Nature (London)* **377,** 348–351.
7. Dhein, J., Walczak, H., Baumler, C., Debatin, K. M. & Krammer, P. H. (1995) *Nature (London)* **373,** 438–441.
8. Hotchkiss, R. S., Swanson, P. E., Knudson, C. M., Chang, K. C., Cobb, J. P., Osborne, D. F., Zollner, K. M., Buchman, T. G., Korsmeyer, S. J. & Karl, I. E. (1999) *J. Immunol.* **162,** 4148–4156.

Organizing Scientific Publications

- ***Goal:*** **find internal categories of publications that share same research areas.**
- **Two sources of interconnections:**
  (1) Words (title, keywords, abstract, body).
  (2) References.
- **Assumptions:**
  - Mixed membership in *K* internal categories.
  - Independent "bag of words" and "bag of references" drawings, conditional on membership scores.

Generative Model

- In our mixed membership model for scientific publications, documents $\mathbf{d} = \left(\left\{ \mathbf{x}_1^{(r_1)} \right\}, \left\{ \mathbf{x}_2^{(r_2)} \right\} \right)$ are generated according to:

$$\lambda \sim \text{Dirichlet}\left(\alpha\right),$$

$$x_1^{(r_1)} \sim \text{Multinomial}\left(p_\lambda\right), \text{ where } p_\lambda = \sum_k \lambda_k \theta_{1k},$$

$$x_2^{(r_2)} \sim \text{Multinomial}\left(q_\lambda\right), \text{ where } q_\lambda = \sum_k \lambda_k \theta_{2k}.$$

- We give distribution to α and then estimate value from data.

PNAS Results

- Fix number of aspects, *K*. For *each* aspect:
  - 39,615 word multinomial parameters,
  - 77,114 reference multinomial parameters,
  - 1 Dirichlet parameter.
- We obtained comparable results from variational approximation and Expectation-Propagation algorithms for 8 aspects.
- Dirichlet parameter estimates for *K*=8:

$$\alpha_1 = 0.0195,\ \alpha_2 = 0.0203, \alpha_3 = 0.0569, \alpha_4 = 0.0346,$$
$$\alpha_5 = 0.0317,\ \alpha_6 = 0.0363, \alpha_7 = 0.0411, \alpha_8 = 0.0255.$$

Posterior Membership Scores

- **Given estimated model parameters, can obtain posterior distribution of article's membership scores via Bayes' theorem (untractable to compute exactly).**
- **Posterior mean membership scores for examples:**

***Ex.1 Reward and punishment*. [Evolution]**

0.0001, **0.9990**, 0.0002, 0.0001, 0.0001, 0.0002, 0.0002, 0.0001

***Ex. 2 Targeted adenovirus-induced expression of IL-10 decreases thymic apoptosis and improves survival in murine sepsis*. [Immunology]**

0.0001, **0.5373**, 0.0002, 0.0001, 0.0001, 0.0001, **0.4619**, 0.0001

Aspect Interpretations

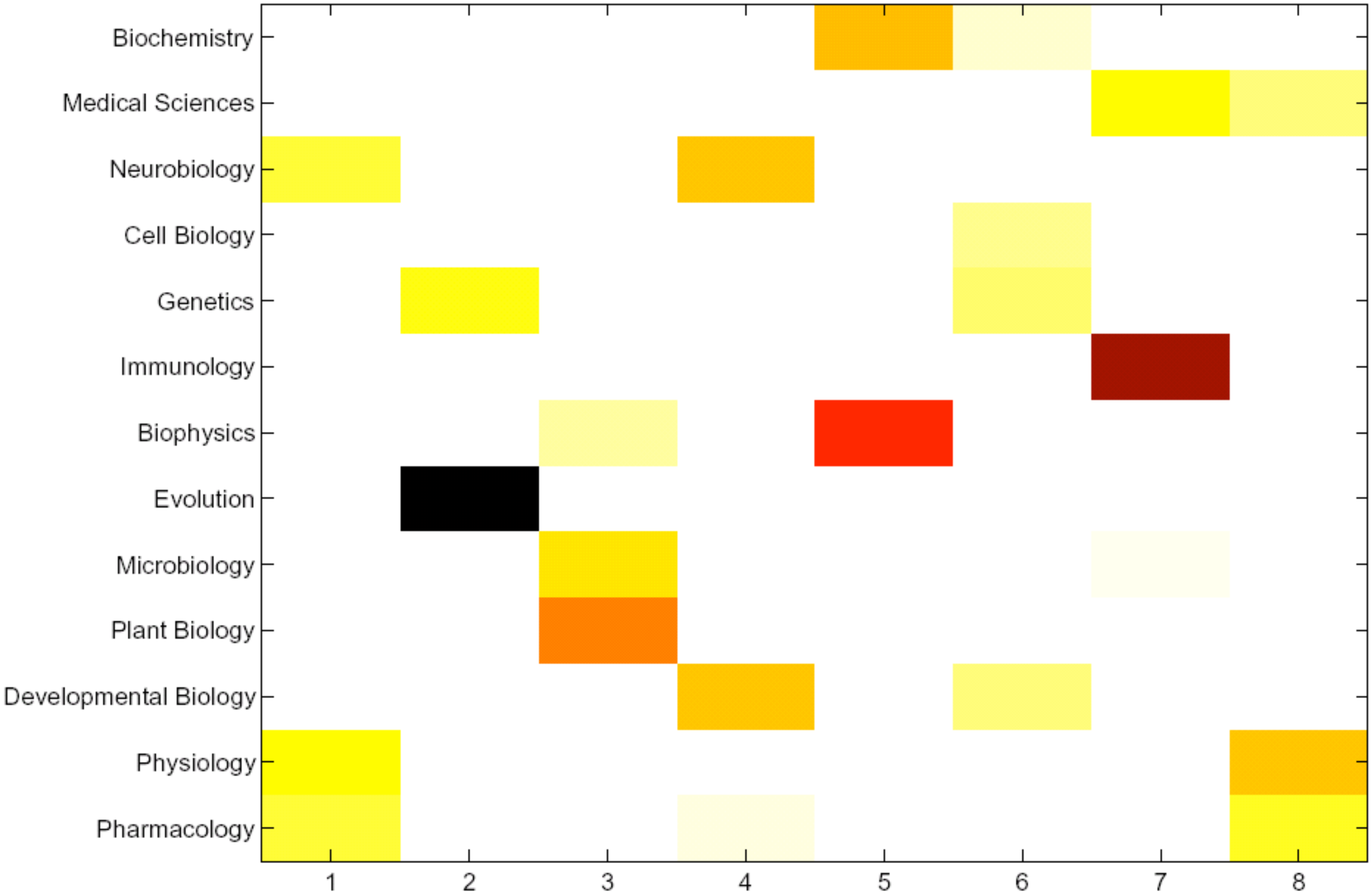
1. Intracellular signal transaction, neurobiology.
2. Evolution, molecular evolution.
3. Plant molecular biology.
4. Developmental biology; brain development.
5. Biochemistry, molecular biology; protein structural biology.
6. Genetics, molecular biology; DNA repair, mutagenesis, cell cycle.
7. Tumor immunology; HIV infection.
8. Endocrinology, reporting of experimental results; molecular mechanisms of obesity.

Mean Decomposition of Loadings

- for 13 highest frequency original classification headings

Topic	1	2	3	4	5	6	7	8
Biochemistry	0.0469	0.0347	0.1810	0.0178	0.3838	0.2057	0.0477	0.0823
Medical Sciences	0.0244	0.0502	0.0938	0.1274	0.0181	0.1075	0.3286	0.2500
Neurobiology	0.2875	0.0398	0.0722	0.3768	0.0196	0.0296	0.0441	0.1304
Cell Biology	0.1691	0.0165	0.1420	0.0684	0.1097	0.2423	0.1637	0.0884
Genetics	0.0141	0.3056	0.1422	0.1532	0.0487	0.2621	0.0395	0.0347
Immunology	0.0127	0.0593	0.1003	0.0413	0.0422	0.0915	0.6244	0.0283
Biophysics	0.0507	0.0295	0.2398	0.0162	0.5496	0.0542	0.0176	0.0423
Evolution	0.0042	0.7679	0.0465	0.0913	0.0289	0.0378	0.0101	0.0133
Microbiology	0.0158	0.1725	0.3431	0.0335	0.0647	0.1174	0.1870	0.0661
Plant Biology	0.1333	0.0983	0.4400	0.0360	0.0462	0.0954	0.0166	0.1344
Developmental Biology	0.0475	0.0288	0.1071	0.3729	0.0274	0.2558	0.0974	0.0631
Physiology	0.3179	0.0275	0.0712	0.1123	0.0258	0.0116	0.0595	0.3743
Pharmacology	0.2883	0.0161	0.0772	0.1965	0.0299	0.0349	0.0537	0.3033

Mean Decomposition of Loadings



Single or Multiple Classification?

	Topic	Total (Dual)	More dual?
1	Biochemistry	2578 (33)	338
2	Medical Sciences	1547 (13)	84
3	Neurobiology	1343 (9)	128
4	Cell Biology	1231 (10)	111
5	Genetics	980 (14)	131
6	Immunology	865 (9)	39
7	Biophysics	636 (40)	62
8	Evolution	510 (12)	167
9	Microbiology	498 (11)	42
10	Plant Biology	488 (4)	54
11	Developmental Biology	366 (2)	43
12	Physiology	340 (1)	34
13	Pharmacology	188 (2)	16
14	Ecology	133 (5)	27
15	Applied Biological Sciences	94 (6)	7
16	Psychology	88 (1)	22
17	Agricultural Sciences	43 (2)	8
18	Population Biology	43 (5)	4
19	Anthropology	10 (0)	2
	Total	11981 (179)	1319

Choosing Number of Aspects

Griffiths and Steyvers (2004) used related version of model on PNAS abstracts only for 1991-2001.

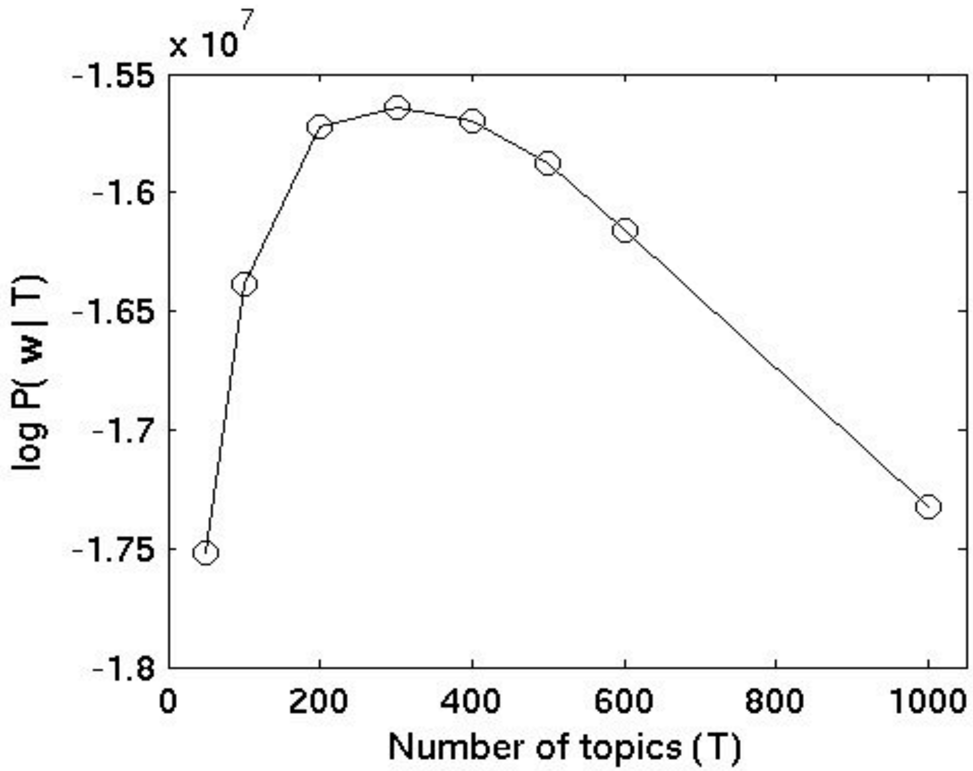
- **Used words from 28,154 abstracts.**
- **20,551 words occurring in at least five abstracts, not on "stop" list.**

Employed Gibbs sampler:

- ***Dirichlet*(α) distribution for membership scores λ;**
- **Fix α at 50/K, where K is the number of aspects;**
- ***Dirichlet*(β) distribution for aspect word probabilities θ;**
- **Fix β at 0.1.**
- **Sample word-aspect assignments and θ.**

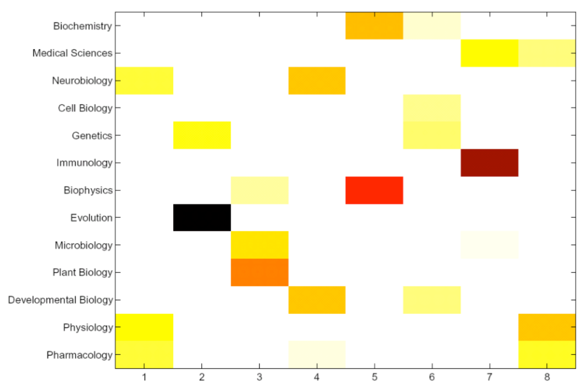
Focus on Number of Aspects

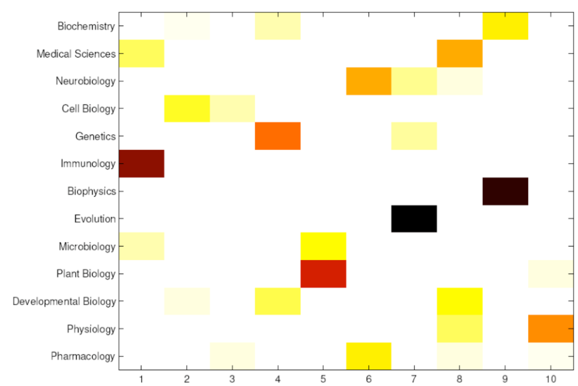
- **Used Pr(data|*K*) for *K* = 50, 100, 200, 300, 400, 500, 600, (1000), integrating out over latent variable, to choose *K* (*T* in their notation).**



Hierarchical Aspect Structure?

- **Aspects function at two different levels in different implementations:**
  - **In word-references model we used *K*=8 and *K*=10 (high level); choice somewhat ad hoc:**





  - **For word model, Griffiths/Steyvers (2004) like *K*=300!**
- **Perhaps we need hierarchical structure for aspects, also with mixed membership.**

Features of Longitudinal Models for NLTCS Data

- **Real panel structure (but irregular spacing because of 1982).**
  - **Disability doesn't simply increase over time:**
    - **Frailty-like or trajectory models.**
    - **Role for marginal modeling????**
    - **Causal models linked to specific illnesses.**
- **Attrition, death, and new entering cohorts.**
  - **Proxy responses are form of informative partial missingness.**
- **Age, Period, and cohort features.**

Features of "Longitudinal" Model for PNAS Articles

- **Syntactic structure.**
- **References and articles have a time stamp!**
  - **Alternative to "bag of references" to reflect time availability of references.**
- **Evolving scientific topics.**

Concluding Remarks

- **Mixed membership approach allows:**
  - **Identification of internal classification categories (unsupervised learning).**
  - **Soft or mixed classifications.**
  - **Combination of characteristics.**
- **Simple idea, complicated estimation:**
  - **Implementation, even in high dimensions.**
- **Challenges remain:**
  - **Full Bayesian calculations; choosing *K*.**
  - **Hierarchical structure for latent categories.**
  - **Modeling longitudinal structure.**

The End

Selected References

- Pritchard, J.K., Stephens M., & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* 155, 945-959.
- Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D. Blei, D. and Jordan, M. (2003) Matching words and pictures. *Journal of Machine Learning Research* 3, 1107–1135.
- Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022.
- Erosheva, E. (2002) *Grade of Membership and Latent Structure Models With Application to Disability Survey Data.* Ph.D. Dissertation, Department of Statistics Carnegie Mellon University.
- Erosheva, E. and Fienberg, S.E. (2005) Bayesian mixed membership models for soft classification. to appear in *Proceedings of the German Classification Society, 2004,* Springer-Verlag.
- Erosheva, E., Fienberg, S.E., and John Lafferty, J. (2004) Mixed-membership models of scientific publications. *PNAS*, 101, 5220-5227.
- Griffiths, T.L, and Steyvers, M. (2004). Finding the topics of science. *PNAS*, 101, 5214-5219.