Mixed Membership Models of Scientific Publications

Elena Erosheva, University of Washington
Stephen Fienberg, Carnegie Mellon University
John Lafferty, Carnegie Mellon University

May 2003

Outline

- Underlying data patterns.
- General class of mixed membership models.
- Interconnectivity of scientific publications.
- Generative model for text and references.
- PNAS database: Biological Sciences.
- Internal categories of publications.
- Next?

Traditional approach: Clustering

How can we learn about underlying data patterns?

- Goal: organize, categorize, summarize.
- Suppose $\mathbf{x}_1,\ldots,\mathbf{x}_J$ are observed variables:
  - disease symptoms,
  - genotypes,
  - weather characteristics;
  - words in text documents.
- Clustering: hierarchical (agglomerative and divisive), non-hierarchical, model-based.

Model-based clustering

- Detect presence of distinct groups.
- Model-based clustering (statistical approach):

  - let $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_J)$ be a sample from some underlying joint distribution $\mathbf{Pr}(\mathbf{x} \mid \theta)$,

  - estimate the joint distribution, indicating presence of groups or lack thereof,

  - each group g is represented by $\mathbf{Pr}_g(\mathbf{x} \mid \theta)$, then

$$\mathbf{Pr}(\mathbf{x} \mid \theta) = \sum_{g=1}^{G} \pi_g \cdot \mathbf{Pr}_g(\mathbf{x} \mid \theta_g).$$

  - mixture model parameters: θ_g, π_g, G.

Traditional mixture models

Each object belongs exclusively to one of the *G* groups or latent classes.

• When attributes have mixed origins from different groups, this assumption may not hold:

- Individual genotypes;
- Individual responses in an attitude survey;
- Words in a scientific article.

In such cases, we say that objects or individuals have mixed membership.

Mixed membership models

- Examples:

Grade of Membership model

(Woodbury, Clive, and Garson, 1978);

Clustering model with admixture for multilocus genotype data

(Pritchard, Stephens, and Donnelly, 2000);

Latent Dirichlet Allocation

(Blei, Ng, and Jordan, 2001).

Assumptions-1

(1) Population level.

- K basis subpopulations;
- $\mathbf{x}_1^{(r)},...,\mathbf{x}_J^{(r)}$ observed response pattern, rth replication;
- subpopulation k is characterized by $f(\mathbf{x}_j \mid \theta_{kj})$;
- responses $\mathbf{x}_1,...,\mathbf{x}_J$ are independent within each subpopulation.

(2) Subject level.

- subjects are characterized by their membership scores, $\lambda = (\lambda_1,...,\lambda_K)$;
- given the membership scores, responses $\mathbf{x}_1,...,\mathbf{x}_J$ are independent;

Assumptions-2

Subject's conditional probabilities:

- Convex combination of corresponding probabilities for *K* subpopulations:

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

Equivalent to a two-stage process:

First stage. Draw latent classification variable z_j:

$$\mathbf{Pr}(\mathbf{z}_j = k \mid \lambda) = \lambda_k.$$

Second stage. Distribution of x_j is determined by the value of latent classification variable, z_j:

$$\mathbf{Pr}(\mathbf{x}_j \mid \mathbf{z}_j = k) = f(\mathbf{x}_j \mid \theta_k).$$

Assumptions-3

(3) Latent variable level.

- Fixed-effects approach:

the membership scores are fixed but unknown.

$$\mathbf{Pr}(\mathbf{x}_j \mid \lambda;\theta) = \sum_k \lambda_k \cdot f(\mathbf{x}_j \mid \theta_{kj}).$$

- Mixed-effects approach:

the membership scores are random, $\lambda \sim \boldsymbol{D}_\alpha$.

$$\mathbf{Pr}(\mathbf{x}_j \mid \alpha;\theta) = \int \left(\sum_k \lambda_k \cdot f(\mathbf{x}_j \mid \theta_{kj}) \right) d\boldsymbol{D}_\alpha(\lambda)$$

Assumptions-4

(4) Sampling scheme:

- J = number of observed distinct characteristics;
- R = number of replications;
- assuming the membership scores are random, the probability is

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

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

Model for scientific publications

- Mixed membership for words and references:
  - membership scores are proportions of document's content originating from each aspect (random);
  - J=2 characteristics (word and reference),
  - R_i replications for words and references, i=1,2.

The PNAS database

- Volumes 94-98 (1997-2001).
- Biological Sciences articles: 92.53% research publications.
- 39,616 unique words in abstracts
- 77,115 unique references in bibliography
- Biological Sciences classifications by PNAS: 19 subtopics.

Biological Sciences Publications

	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

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:
  - *K* internal categories;
  - mixed membership in *K* internal categories;
  - "bag of words" and "bag of references"

  drawings, conditional on the 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 the following:

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

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

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

Estimation

- Bayesian approach: obtain samples from posterior distributions with Markov chain Monte Carlo methods.
- Direct approximations:
  - Variational approximation algorithm.
  - Expectation-Propagation algorithm.
- Obtained comparable results from variational approximation and Expectation-Propagation algorithms for 8 aspects.

Estimation

- Assume K aspects.
- For each aspect:
  - 39,615 word multinomial parameters,
  - 77,114 reference multinomial parameters,
  - 1 Dirichlet parameter.
- Dirichlet parameter estimates, K=8:

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

High probability words by aspect

Aspect 1		Aspect 2		Aspect 3		Aspect 4	
ca2+	0.0062	species	0.0040	sequence	0.0024	development	0.0034
channel	0.0047	sequence	0.0026	acid	0.0020	neurons	0.0034
membrane	0.0047	sequences	0.0024	plants	0.0018	brain	0.0029
channels	0.0040	genetic	0.0024	cdna	0.0017	mouse	0.0025
receptors	0.0028	genome	0.0022	mutant	0.0015	normal	0.0024
synaptic	0.0026	evolution	0.0020	single	0.0015	expressed	0.0021
neurons	0.0022	among	0.0017	enzyme	0.0015	cortex	0.0019
g	0.0021	population	0.0016	plant	0.0014	embryonic	0.0017
calcium	0.0021	most	0.0016	identified	0.0013	adult	0.0017
activation	0.0020	chromosome	0.0015	amino	0.0013	neuronal	0.0016
release	0.0020	selection	0.0015	expressed	0.0013	function	0.0016
kinase	0.0019	populations	0.0014	mutants	0.0013	neural	0.0015
subunit	0.0019	three	0.0014	molecules	0.0012	early	0.0014
intracellular	0.0017	based	0.0013	based	0.0012	patients	0.0014
acid	0.0016	variation	0.0013	kda	0.0011	functional	0.0013

High probability words by aspect

Aspect 5		Aspect 6		Aspect 7		Aspect 8	
residues	0.0028	transcription	0.0060	il	0.0046	increased	0.0027
enzyme	0.0023	nuclear	0.0036	tumor	0.0040	receptors	0.0023
active	0.0020	promoter	0.0031	activation	0.0036	g	0.0022
terminal	0.0019	transcriptional	0.0030	hiv	0.0032	p	0.0022
amino	0.0019	p53	0.0029	apoptosis	0.0031	insulin	0.0018
rna	0.0018	rna	0.0027	kinase	0.0028	effects	0.0018
structural	0.0018	kinase	0.0024	antigen	0.0026	increase	0.0018
state	0.0018	yeast	0.0024	virus	0.0025	acid	0.0018
folding	0.0017	function	0.0022	gamma	0.0021	effect	0.0016
sequence	0.0017	activation	0.0020	infection	0.0021	fold	0.0016
form	0.0016	sequence	0.0018	immune	0.0020	reduced	0.0016
peptide	0.0016	terminal	0.0018	signaling	0.0018	treatment	0.0016
atp	0.0015	cycle	0.0018	death	0.0017	glucose	0.0016
helix	0.0015	mutations	0.0017	activated	0.0017	mrna	0.0015
substrate	0.0015	factors	0.0017	vivo	0.0017	rats	0.0015

High probability references

Aspect 1

Author	Journal, Year	C
HAMILL OP	PFLUG ARCH EUR J PHY, 1981	72
LAEMMLI UK	Nature, 1970	322
HILLE B	IONIC CHANNELS EXCIT, 1992	58
BLISS TVP	NATURE, 1993	54
SUDHOF TC	NATURE, 1995	33
GRYNKIEWICZ G	J BIOL CHEM, 1985	31
SAMBROOK J	MOL CLONING LAB MANU, 1989	764
SHERRINGTON R	NATURE, 1995	33
ROTHMAN JE	NATURE, 1994	27
SIMONS K	NATURE, 1997	35
SOLLNER T	NATURE, 1993	25
ROTHMAN JE	SCIENCE, 1996	24
THINAKARAN G	NEURON, 1996	23
TOWBIN H	P NATL ACAD SCI USA, 1979	86
BERMAN DM	CELL, 1996	21

High probability references

Aspect 4

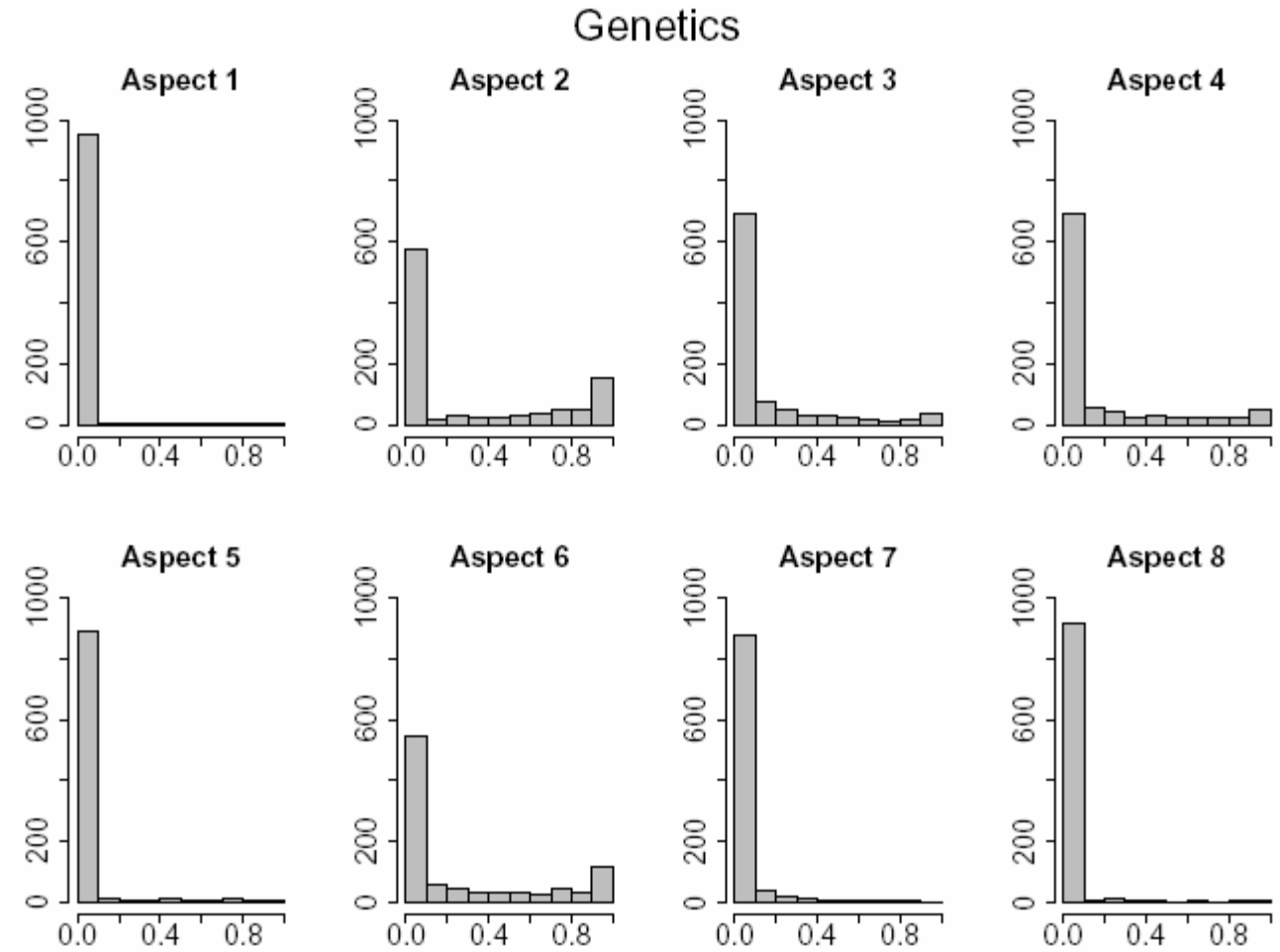
Author	Journal, Year	C
HOGAN B	MANIPULATING MOUSE E, 1994	68
CHOMCZYNSKI P	ANAL BIOCHEM, 1987	206
TALAIRACH J	COPLANAR STEREOTAXIC, 1988	60
PAXINOS G	RAT BRAIN STEREOTAXI, 1986	38
SAMBROOK J	MOL CLONING LAB MANU, 1989	764
NAGY A	P NATL ACAD SCI USA, 1993	39
MANSOUR SL	NATURE, 1988	37
BRAND AH	DEVELOPMENT, 1993	46
HOGAN B	MANIPULATING MOUSE E, 1986	32
TYBULEWICZ VLJ	CELL, 1991	46
KWONG KK	P NATL ACAD SCI USA, 1992	24
DUNLAP JC	CELL, 1999	19
LI E	CELL, 1992	35
ALTSCHUL SF	J MOL BIOL, 1990	253
EISEN MB	P NATL ACAD SCI USA, 1998	60

High probability references

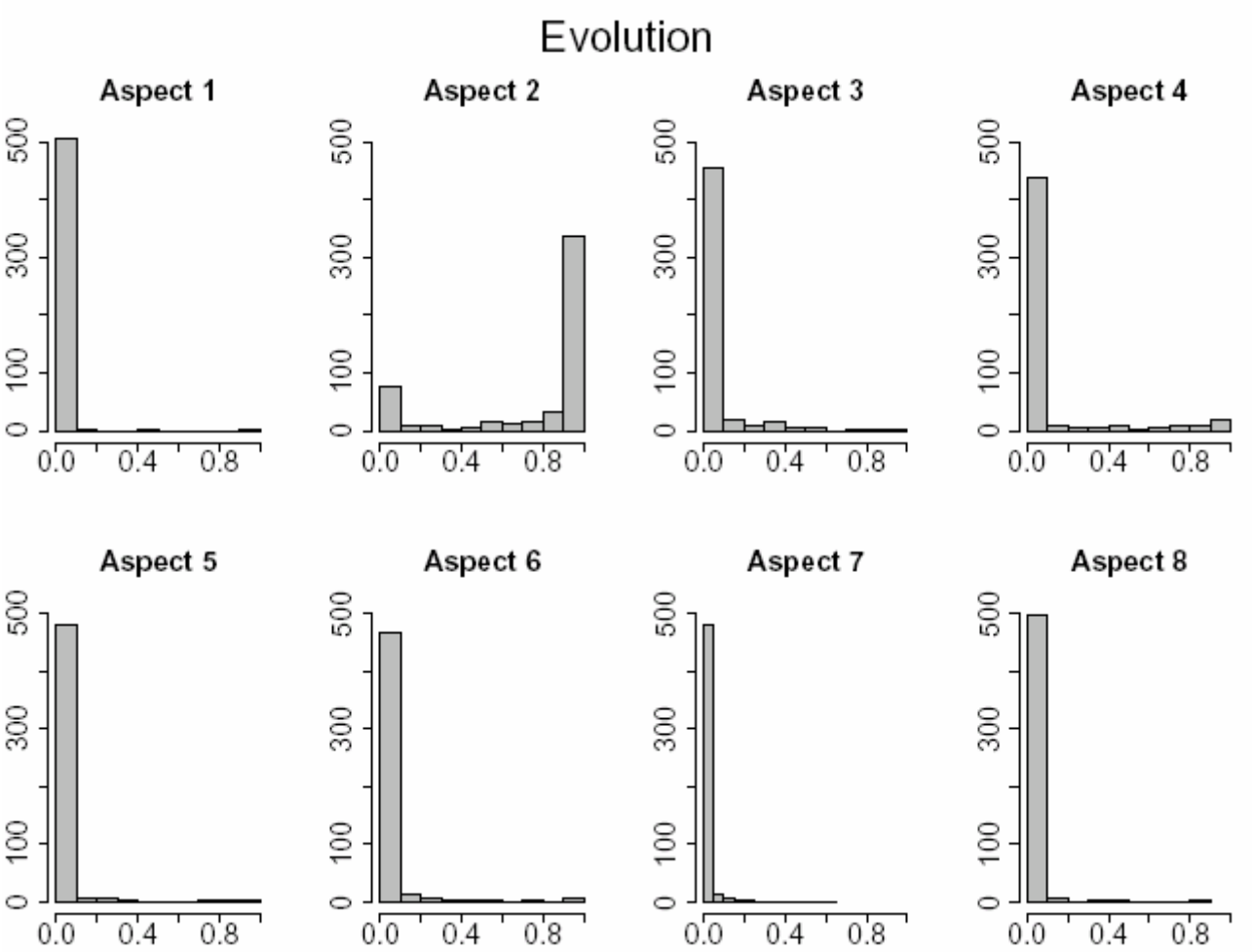
Aspect 7

Author	Journal, Year	C
DENG HK	NATURE, 1996	46
DRAGIC T	NATURE, 1996	45
DORANZ BJ	CELL, 1996	45
FENG Y	SCIENCE, 1996	43
ALKHATIB G	SCIENCE, 1996	43
COCCHI F	SCIENCE, 1995	41
CHOE H	CELL, 1996	41
THOMPSON CB	SCIENCE, 1995	38
ZOU H	CELL, 1997	40
DARNELL JE	SCIENCE, 1994	40
MUZIO M	CELL, 1996	35
LI P	CELL, 1997	36
XIA ZG	SCIENCE, 1995	38
BOLDIN MP	CELL, 1996	34
PEAR WS	P NATL ACAD SCI USA 1993	57

Distribution of membership scores

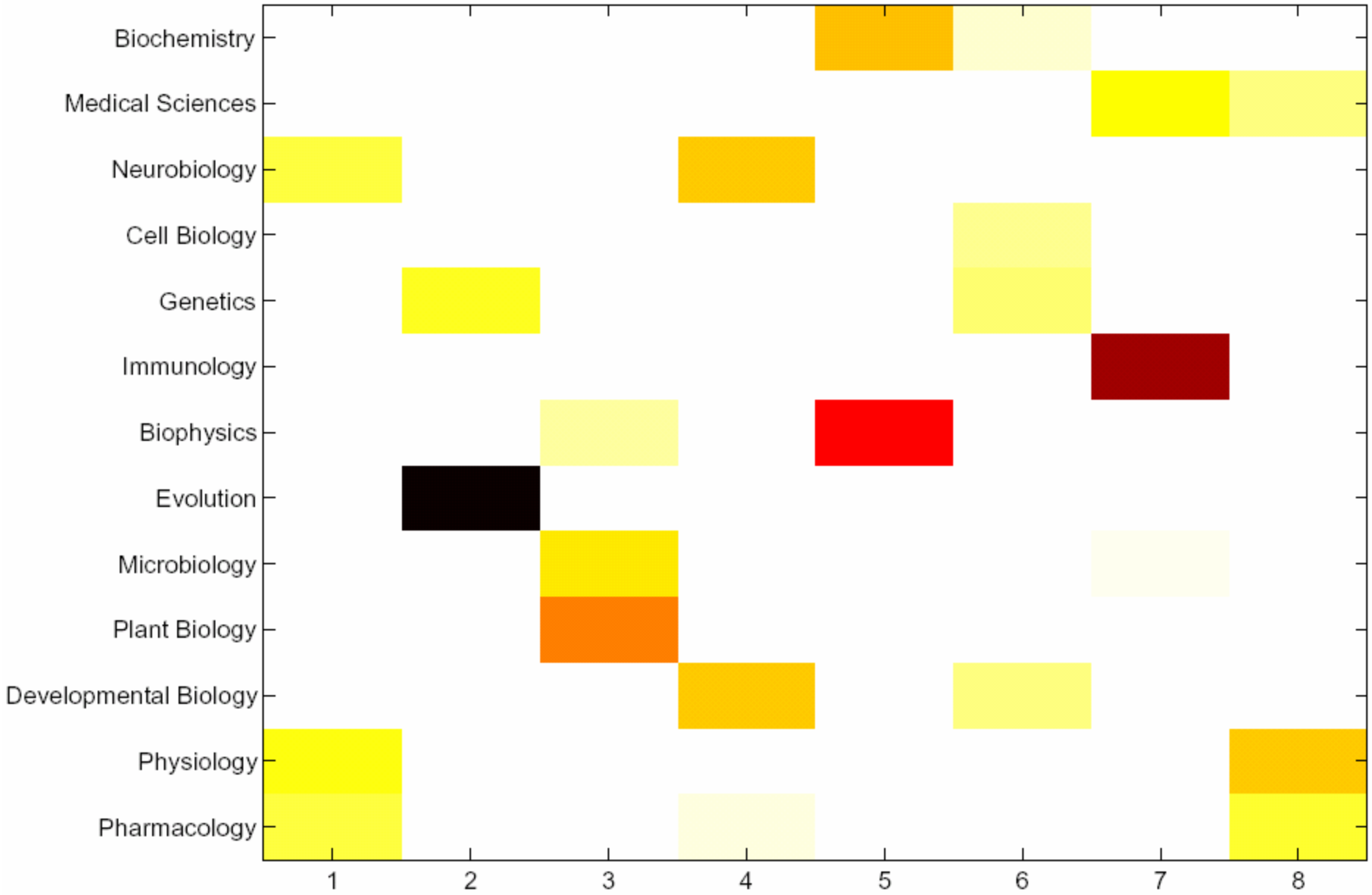


Distribution of membership scores



Mean decompositions of loadings

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


Biochemistry
Medical Sciences
Neurobiology
Cell Biology
Genetics
Immunology
Biophysics
Evolution
Microbiology
Plant Biology
Developmental Biology
Physiology
Pharmacology
1
2
3
4
5
6
7
8

May 2003

Role of statistical models

- Simplifying assumptions:
  - do not account for syntax and grammar,
  - do not account for types of references,
  - “bag of words” and “bag of references”.
- Despite oversimplifying assumptions, models are able to pick up some of the fine structure.
- The model for words has been informative.
- “All models are wrong, but some are useful.” Box (1976)

Concluding remarks

- Allows to identify internal categories of publications.
- Allows to combine distinct characteristics such as words and references.
- Simple idea, complicated estimation.
- PNAS database: most traditional discipline classifications in Biological Sciences have mixed distributions over internal categories.

Selected references

- Woodbury, M.A., Clive, J., & Garson, A. (1978) Mathematical typology: A Grade of Membership technique for obtaining disease definition. *Computers and Biomedical Research* 11, 277-298.
- Pritchard, J.K., Stephens M., & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* 155, 945-959.
- Blei, D.M., Ng, A.Y., & Jordan, M.I. (2001) Latent Dirichlet allocation. *Advances in Neural Information Processing Systems*.
- Cohn, D., and Hofmann, T. (2001) The missing link - a probabilistic model of document content and hypertext connectivity. *Neural Information Processing Systems 13.*
- 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.