The Aspect Bernoulli model: multiple causes of presences and absences

Ella Bingham

ella@iki.fi

HIIT BRU

ntroduction: Latent variable models - why

- latent = hidden, unobserved
- latent variable models $\approx$ multiple cause models, mixture models, factor models, . . .

ntroduction: Latent variable models - why

- latent = hidden, unobserved
- latent variable models $\approx$ multiple cause models, mixture models, factor models, . . .
- A small number of unknown (latent) variables combine to explain a large data set
- A natural framework for unifying statistical inference and clustering

Aspect Bernoulli (AB)

- Joint work with Dr Ata Kabán
- A probabilistic multiple cause model for 0-1 data

Aspect Bernoulli (AB)

- Joint work with Dr Ata Kabán
- A probabilistic multiple cause model for 0-1 data
- Novelty: some components explicitely account for noise
- Suppose a separate noise process has turned some 0s to 1, and vice versa
- Detect the noise factors, and correct the data accordingly

Aspect Bernoulli (AB)

- Joint work with Dr Ata Kabán
- A probabilistic multiple cause model for 0-1 data
- Novelty: some components explicitely account for noise
- Suppose a separate noise process has turned some 0s to 1, and vice versa
- Detect the noise factors, and correct the data accordingly
- Related methods: Mixtures of Bernoulli,
  Probabilistic Latent Semantic Analysis “aspect Multinomial”,
  Latent Dirichlet Allocation, Multinomial PCA
  Logistic PCA

- AB detects and distinguishes between “true absences” and “missing presences” (both of which coded as 0)
- Examples: why a word is not present in a document; why a mammal species is not found at a site of excavation

- AB detects and distinguishes between “true absences” and “missing presences” (both of which coded as 0)
- Examples: why a word is not present in a document; why a mammal species is not found at a site of excavation
- Similarly, between “true presences” and “added presences” (both of which coded as 1)
- Example: extra black pixels added to an image

- Notation:
  $n = 1, \ldots, N$ observations
  $t = 1, \ldots, T$ attributes
  $k = 1, \ldots, K$ latent aspects (=mixture components = factors)
  $\mathbf{x}_n$ T-dimensional observation;
  x_{tn} its value at attribute t

- Notation:
  $n = 1, \ldots, N$ observations
  $t = 1, \ldots, T$ attributes
  $k = 1, \ldots, K$ latent aspects (=mixture components = factors)
  $\mathbf{x}_n$ T-dimensional observation;
  x_{tn} its value at attribute t
- Data generation process:
- $(s_{1n}, \ldots, s_{Kn})$ is a distribution of latent aspects k for observation n

- Notation:
  $n = 1, \ldots, N$ observations
  $t = 1, \ldots, T$ attributes
  $k = 1, \ldots, K$ latent aspects (=mixture components = factors)
  $\mathbf{x}_n$ T-dimensional observation;
  x_{tn} its value at attribute t
- Data generation process:
- $(s_{1n}, \ldots, s_{Kn})$ is a distribution of latent aspects k for observation n
- At each attribute t of observation n, pick a component k from this distribution with probability $s_{kn} = prob(k|n)$

- Notation:
  $n = 1, \ldots, N$ observations
  $t = 1, \ldots, T$ attributes
  $k = 1, \ldots, K$ latent aspects (=mixture components = factors)
  $\mathbf{x}_n$ T-dimensional observation;
  x_{tn} its value at attribute t
- Data generation process:
- $(s_{1n}, \ldots, s_{Kn})$ is a distribution of latent aspects k for observation n
- At each attribute t of observation n, pick a component k from this distribution with probability $s_{kn} = prob(k|n)$
- Then generate 1 or 0 with $a_{tk} = prob(1|t,k)$

- Likelihood:

$$p(\mathbf{x}_n|model) = \prod_t \sum_k s_{kn} a_{tk}^{x_{tn}} (1 - a_{tk})^{1-x_{tn}}$$
$$= \prod_t (\sum_k a_{tk} s_{kn})^{x_{tn}} (1 - \sum_k a_{tk} s_{kn})^{1-x_{tn}}$$

- Likelihood:

$$p(\mathbf{x}_n|model) = \prod_t \sum_k s_{kn} a_{tk}^{x_{tn}} (1 - a_{tk})^{1-x_{tn}}$$
$$= \prod_t (\sum_k a_{tk} s_{kn})^{x_{tn}} (1 - \sum_k a_{tk} s_{kn})^{1-x_{tn}}$$

- Decomposes the Bernoulli mean!
- Model estimated by EM algorithm

- Likelihood:

$$p(\mathbf{x}_n|model) = \prod_t \sum_k s_{kn} a_{tk}^{x_{tn}} (1-a_{tk})^{1-x_{tn}}$$
$$= \prod_t (\sum_k a_{tk} s_{kn})^{x_{tn}} (1 - \sum_k a_{tk} s_{kn})^{1-x_{tn}}$$

- Decomposes the Bernoulli mean!
- Model estimated by EM algorithm
- Noise is automatically factored into “phantom” components:
  $1 \longrightarrow 0$ noise is modeled by a component k having
  $a_{tk} = prob(1|t,k) \approx 0$ at all t
  $0 \longrightarrow 1$ noise is modeled by $a_{tk} \approx 1$ at all t

Experimental results

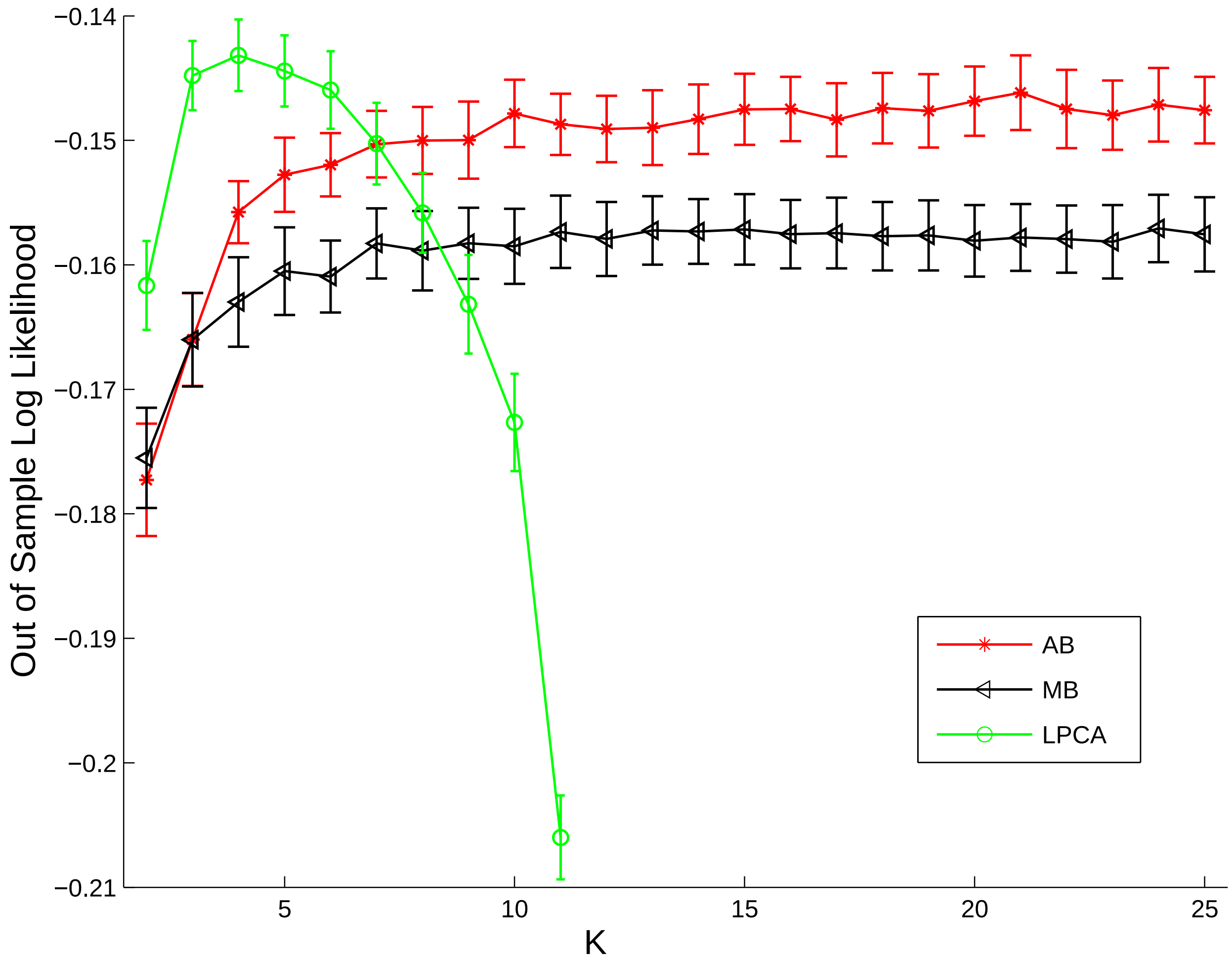
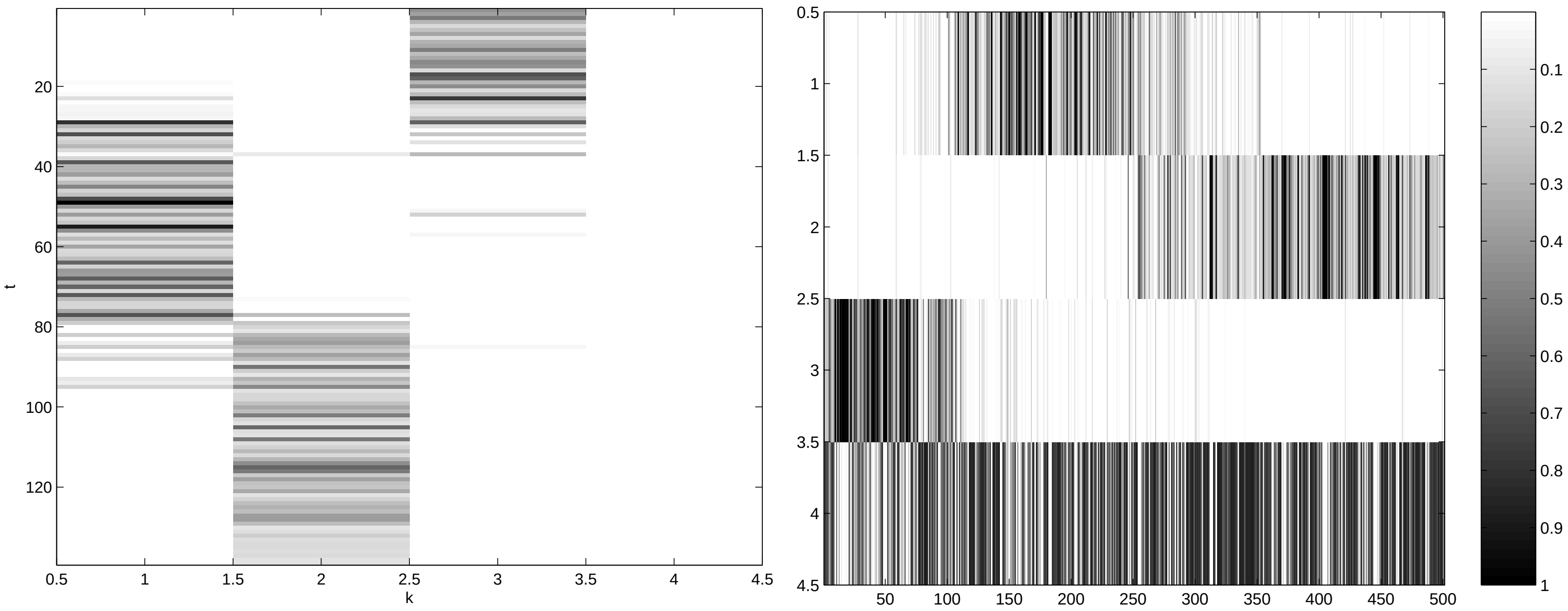


Figure 1: Out of sample log likelihood in paleo data

Parameters in paleo data

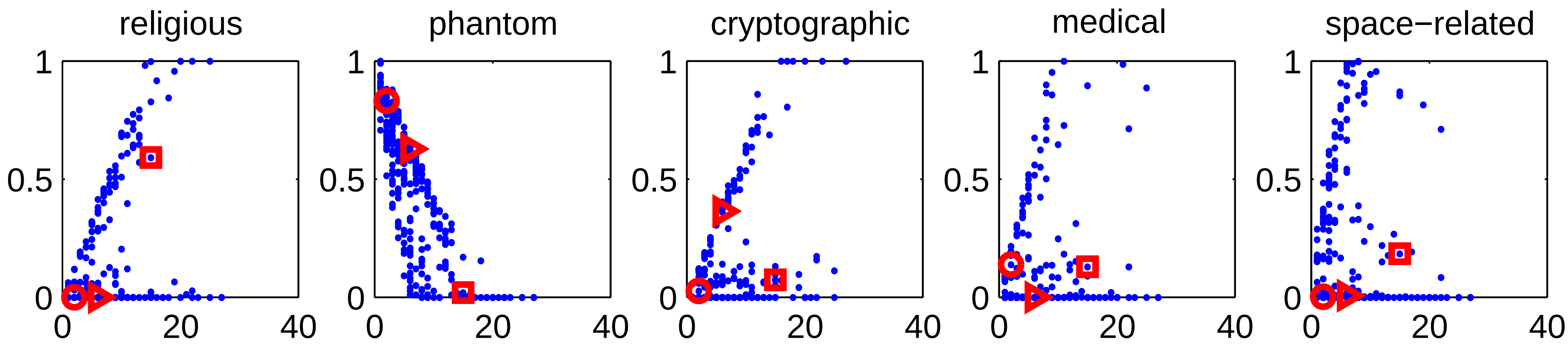


Left: $a_{tk} = P(1|k,t)$ at mammal genera t.

Right: $s_{kn} = P(k|n)$ at sites of excavation n

Topics in 4 Newsgroups

religious	*phantom*	*cryptographic*	*medical*	*space-related*
god 1.00	agre 1.3e-03	kei 1.00	effect 0.84	space 0.76
christian 1.00	sternlight 1.0e-11	encrypt 1.00	peopl 0.72	nasa 0.59
peopl 0.95	bless 3.2e-12	**system** 1.00	medic 0.66	orbit 0.49
rutger 0.81	truth 2.5e-15	govern 0.90	doctor 0.52	man 0.37
word 0.63	peopl 2.4e-15	public 0.89	patient 0.47	cost 0.35
church 0.63	comput 2.8e-16	clipper 0.84	diseas 0.42	**system** 0.34
bibl 0.61	system 8.6e-19	chip 0.83	treatment 0.40	pat 0.33
faith 0.60	man 1.1e-19	secur 0.82	medicin 0.40	launch 0.32
christ 0.59	nsa 1.0e-21	peopl 0.70	food 0.35	mission 0.30
jesu 0.56	shuttl 4.1e-22	comput 0.65	med 0.33	flight 0.28



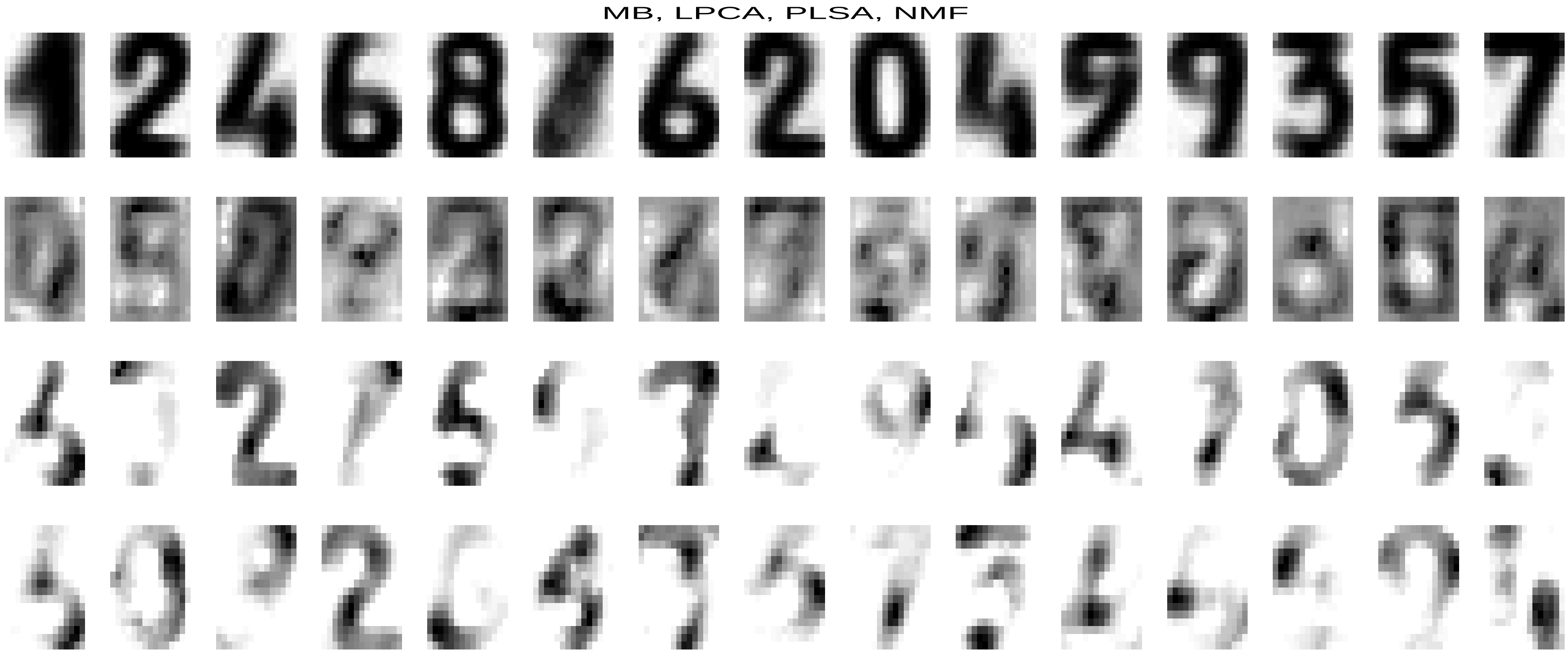
$s_{kn} = P(k|n)$ vs number of words in document n. ○: 'system' 'medicin'; □: 'peopl' 'public' 'system' 'agre' 'faith' 'accept' 'christ' 'teach' 'clinic' 'mission' 'religion' 'jesu' 'holi' 'doctrin' 'scriptur'; ▷: 'govern' 'secur' 'access' 'scheme' 'system' 'devic'

“Query expansion”

govern secur access scheme system devic	kei 0.99 encrypt 0.99 public 0.98 clipper 0.92 chip 0.91 peopl 0.89 comput 0.84 escro
encrypt decrypt tap	system 1.00 kei 1.00 public 1.00 govern 0.98 secur 0.98 clipper 0.97 chip 0.97 peopl (
algorithm encrypt secur access peopl scheme system comput	kei 0.98 public 0.97 govern 0.92 clipper 0.87 chip 0.85 escrow 0.75 secret 0.63 nsa 0.
peopl effect diseas medicin diagnos	medic 0.98 doctor 0.77 patient 0.75 treatment 0.71 physician 0.66 food 0.66 symptom
system medicin	effect 0.97 medic 0.96 peopl 0.96 doctor 0.92 patient 0.92 diseas 0.91 treatment 0.91
peopl secret effect cost doctor patient food pain	medic 0.48 diseas 0.28 treatment 0.27 medicin 0.27 physician 0.24 symptom 0.24 me

Corrupted handwritten digits

Other methods



MB, LPCA, PLSA, NMF

Conclusion

- Multiple cause model for 0-1 data
- More expressive power than in Bernoulli mixtures
- Parameters easy to interpret
- Noise explicitely factored into separate components
- Ongoing work