

# ICA-based Binary Feature Construction

**Ata Kabán**

University of Birmingham, UK

**Ella Bingham**

HIIT BRU, University of Helsinki

## Motivation

- ▶ Abundant sources of binary data
  - digital b/w images
  - digital text repositories
  - paleo-ecological recordings
  - DNA fingerprints
  
- ▶ Need appropriate tools for processing binary data
  - exploratory analysis
  - denoising
  - feature construction
  
- ▶ ICA - a useful statistical data representation principle
  - little work exists for the binary case

## Model

$\mathbf{x}_1, \dots, \mathbf{x}_N$  multivariate binary data.

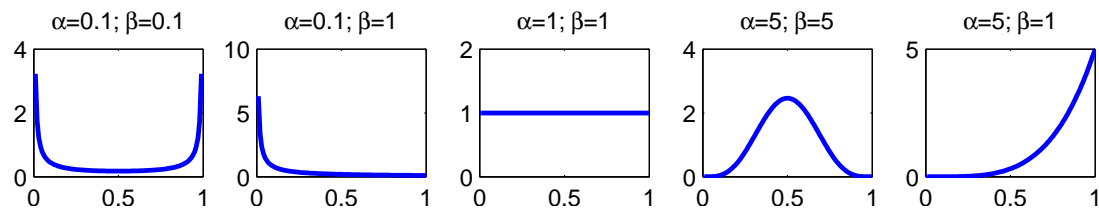
### ► Bernoulli likelihood

$$P(\mathbf{x}_n | \mathbf{b}) = \prod_t \left( \sum_k a_{tk} b_k \right)^{x_{tn}} \left( 1 - \sum_k a_{tk} b_k \right)^{1-x_{tn}} \quad (1)$$

### ► Independent beta sources

$$p(\mathbf{b}) = \prod_k p(b_k) = \prod_k B(b_k | \alpha_k^0, \beta_k^0) db_k \quad (2)$$

where  $B(b | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} (1 - b)^{\beta-1} b^{\alpha-1}$  is the Beta density.



## Model (cont'd)

So,

$$P(\mathbf{x}_n) = \int \prod_t \left( \sum_k a_{tk} b_k \right)^{x_{tn}} \left( 1 - \sum_k a_{tk} b_k \right)^{1-x_{tn}} \prod_k B(b_k | \alpha_k^0, \beta_k^0) db_k$$

- The mixing through  $a_{tk}$  is assumed to be convex.

$$\sum_k a_{tk} = 1, a_{tk} \geq 0, \forall t, k \quad (3)$$

- this does not restrict the number of components being mixed
- computationally convenient — it follows that

$$\sum_k a_{tk} b_k \in [0, 1] \quad (4)$$

is already a probability (no need for a ‘link’ function)

## Variational inference & estimation

- Integration intractable
- Create lower bound using Jensen's inequality

$$\log P(\mathbf{x}_n) = \log \int P(\mathbf{x}_n | \mathbf{b}) \prod_k B(b_k | \alpha_k^0, \beta_k^0) db_k \quad (5)$$

$$\geq \int \prod_k q_n(b_k) \log \frac{\prod_t P(x_{tn} | \mathbf{b}) \prod_k B(b_k | \alpha_k^0, \beta_k^0)}{\prod_k q_n(b_k)} db_k \quad (6)$$

where  $\prod_k q_n(b_k)$  is a factorial variational posterior.

- ...still intractable due to the likelihood term

- ...further lower bound this term, using the convexity of mixing

$$\begin{aligned} \log P(x_{tn}|\mathbf{b}) &= \log \left\{ \left( \sum_k a_{tk} b_k \right)^{x_{tn}} \left( 1 - \sum_k a_{tk} b_k \right)^{1-x_{tn}} \right\} \\ &= \log \left\{ \sum_k a_{tk} b_k^{x_{tn}} (1 - b_k)^{1-x_{tn}} \right\} \geq \sum_k Q_{k|t,n,x_{tn}} \log \frac{a_{tk} b_k^{x_{tn}} (1 - b_k)^{1-x_{tn}}}{Q_{k|t,n,x_{tn}}} \quad (7) \end{aligned}$$

Here  $Q_{k|t,n,x_{tn}} \geq 0$ ,  $\sum_k Q_{k|t,n,x_{tn}} = 1$  is a discrete variational distribution with values in  $\{1, ..K\}$ , where  $K$  denotes the number of components.

- The variational posteriors:

- Let  $q_n(b_k) = B(b_k|\alpha_{kn}, \beta_{kn})$  be parameterised Beta variational posteriors with variational parameters  $\alpha_{kn}, \beta_{kn}$ .
- $Q_{k|t,n,x_{tn}}$  is a discrete variational posterior

## Variational E-step

$$Q_{k|t,n,x_{tn}} \propto a_{tk} (e^{\langle \log b_{kn} \rangle})^{x_{tn}} (e^{\langle \log(1-b_{kn}) \rangle})^{1-x_{tn}} \quad (8)$$

$$\alpha_{kn} = \alpha_k^0 + \sum_t x_{tn} Q_{k|t,n,(x_{tn}=0)} \quad (9)$$

$$\beta_{kn} = \beta_k^0 + \sum_t (1 - x_{tn}) Q_{k|t,n,(x_{tn}=1)} \quad (10)$$

where required variational posterior expectations are easily evaluated as  $\langle \log b_{kn} \rangle \equiv E_{q_n(b_k)}[\log b_k] = \psi(\alpha_{kn}) - \psi(\alpha_{kn} + \beta_{kn})$  and  $\langle \log(1 - b_{kn}) \rangle \equiv E_{q_n(b_k)}[\log(1 - b_k)] = \psi(\beta_{kn}) - \psi(\alpha_{kn} + \beta_{kn})$ .

- Observe  $Q_{k|t,n,x_{tn}}$  need not stored, instead can be replaced into (9) and (10)

## Algorithm

### ► Var E-step

$$\alpha_{kn} = \alpha_k^0 + e^{\langle \log b_{kn} \rangle} \sum_t \frac{x_{tn} a_{tk}}{\sum_k a_{tk} e^{\langle \log b_{kn} \rangle}} \quad (11)$$

$$\beta_{kn} = \beta_k^0 + e^{\langle \log(1-b_{kn}) \rangle} \sum_t \frac{(1-x_{tn}) a_{tk}}{\sum_k a_{tk} e^{\langle \log(1-b_{kn}) \rangle}} \quad (12)$$

### ► M-step

$$a_{tk} \propto a_{tk} \left\{ \sum_n \frac{x_{tn}}{\sum_k a_{tk} e^{\langle \log b_{kn} \rangle}} e^{\langle \log b_{kn} \rangle} + \frac{1-x_{tn}}{\sum_k a_{tk} e^{\langle \log(1-b_{kn}) \rangle}} e^{\langle \log(1-b_{kn}) \rangle} \right\} \quad (13)$$

## Model complexity control

- Prior on the mixing coefficients

$$p(\mathbf{a}) = \text{Dirichlet}(\mathbf{a}|\gamma^0) \quad (14)$$

Nr of components is the maximiser of the evidence bound

$$E_{q_t(\mathbf{a})}[\mathcal{L}^{\text{bound}}] + E_{q_t(\mathbf{a})}[\log \text{Dir}(\mathbf{a}|\gamma^0)] - E_{q_t(\mathbf{a})}[\log q_t(\mathbf{a})] \quad (15)$$

where  $q_t(\mathbf{a}) = \text{Dir}(\mathbf{a}|\gamma_t)$  variational posterior of  $\mathbf{a}$

- Variational E-step as before, but  $a_{tk}$  replaced by  $e^{\langle \log \gamma_{tk} \rangle}$
- Variational M-step

$$\gamma_{tk} = \gamma_k^0 + e^{\langle \log a_{tk} \rangle} \left\{ \sum_n \frac{x_{tn} e^{\langle \log b_{kn} \rangle}}{\sum_k e^{\langle \log a_{tk} \rangle} e^{\langle \log b_{kn} \rangle}} + \frac{(1 - x_{tn}) e^{\langle \log(1-b_{kn}) \rangle}}{\sum_k e^{\langle \log a_{tk} \rangle} e^{\langle \log(1-b_{kn}) \rangle}} \right\}$$

where  $\langle \log a_{tk} \rangle \equiv E_{q_t(\mathbf{a})}[\log a_k] = \psi(\gamma_{tk}) - \psi(\sum_{k'} \gamma_{tk'})$

## Analyst input

- ▶ Independent components may be easier to comprehend separately
- ▶ For each IC  $\langle \mathbf{b}_k \rangle$ , analyst provides a probability  $P(u|k)$  expressing relevance. Noise components may be given  $P(u|k) = 0$ .
- ▶ Post-processing the posterior expectations returned by the unsupervised algorithm by inverting the user feedback: Denote by  $P_t(k)$  the posterior expectations  $\langle a_{tk} \rangle$ . By Bayes rule,  
$$\langle a_{tk} \rangle_{postproc} := P_t(k|u) \propto P_t(k)P(u|k)$$
- ▶ Let  $q_t(\mathbf{a}|\mathbf{u})$  be the Dirichlet whose expectation is  $\langle a_{tk}|\mathbf{u} \rangle = \langle a_{tk} \rangle_{postproc}$
- ▶ The posterior reconstruction of the data will be  
$$\langle \hat{x}_{tn}|\mathbf{u} \rangle = p(\hat{x}_{tn} = 1|\mathbf{X}, \mathbf{u}) = \sum_k \langle a_{tk}|\mathbf{u} \rangle \langle b_{kn} \rangle$$

## Restoration of corrupted b/w images

- ▶ 1000 images of handwritten digits (15 × 16 pixels each)
- ▶ artificially corrupted by simulating a uniformly varying process of degradation, which turns off some of the pixels
- ▶ How can we detect the corruption automatically?
- ▶ How can we restore the originals?

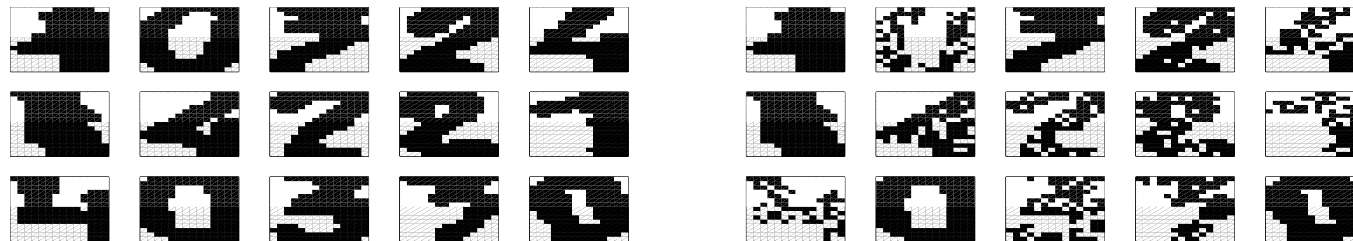


Figure 1: Examples of clean (left) and corrupted (right) images.

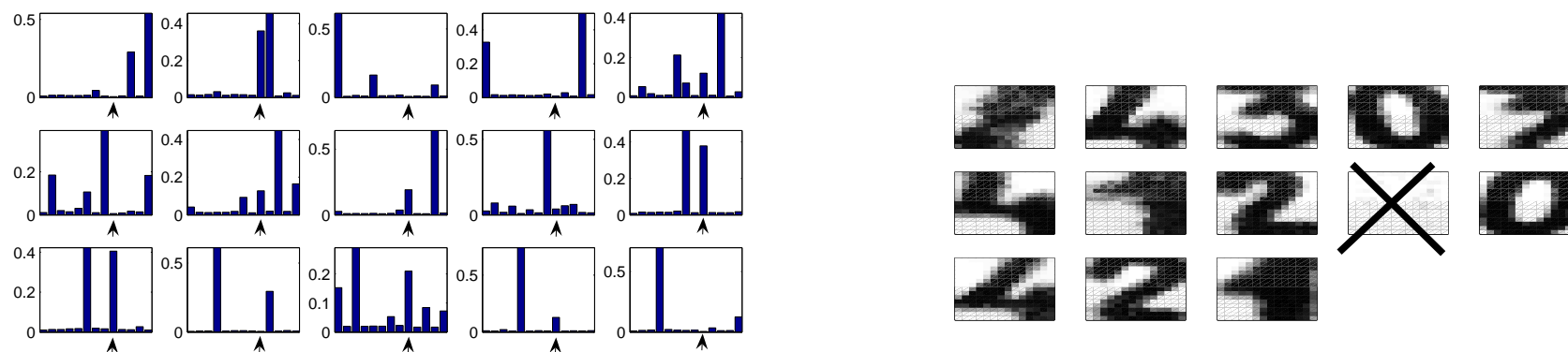


Figure 2: Right: ICs of the corrupted images; Left: Mixing coefficients for the data shown earlier. Small arrow heads point to the mixing coefficients associated with the white component.

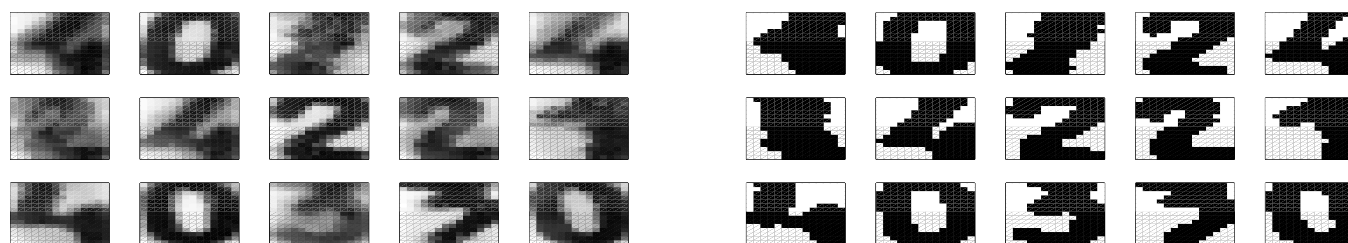


Figure 3: Reconstructed grey-scale (left) and binary (right) images after the post-processing.

- None of the existing binary data analysis methods tested, except our binary ICA was able to separate out the noise factor.
- Therefore analyst feedback based correction is not applicable to the other methods.

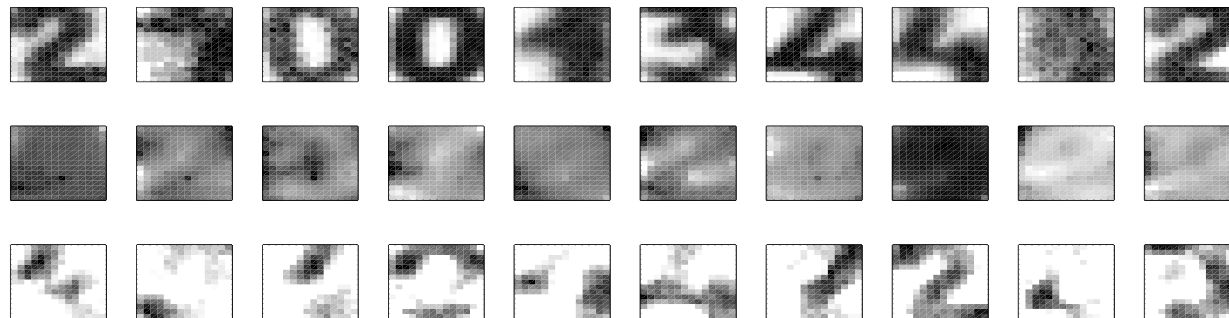


Figure 4: Example representation bases created on artificially corrupted binary handwritten digit images by MB (top row), LPCA (middle row) and binary NMF (last row) respectively.

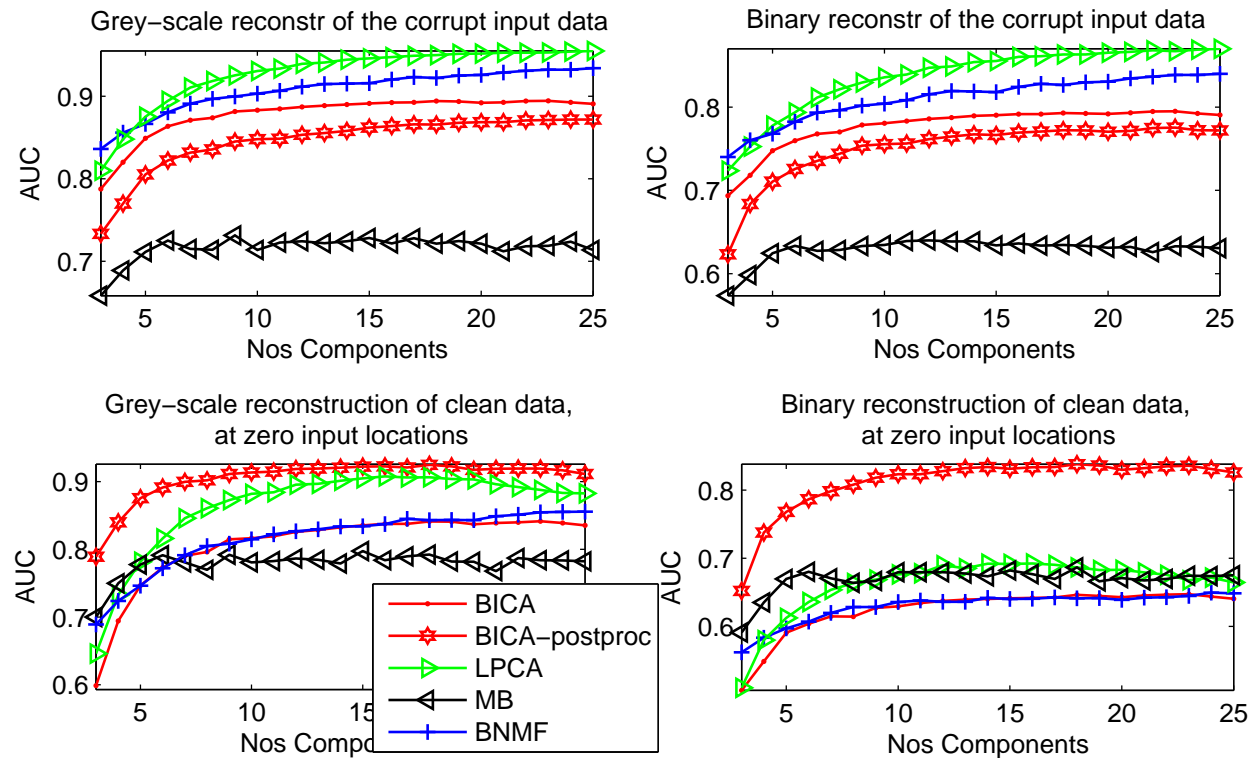


Figure 5: BICA with user feedback postprocessing compared to: unsupervised BICA, Logistic PCA (Tipping '99), Mixture of Bernoulli, and a binary version of NMF. LPCA

## Independent components of binary coded text

- Usenet messages from 4 different topics of discussion.

religious	'blank'	cryptographic	medical	space-related
god 1.00	0.01	<b>system 1.00</b>	effect 1.00	space 0.76
christ 1.00	0.00	kei 1.00	medic 0.99	nasa 0.61
peopl 0.99	0.00	encrypt 1.00	peopl 0.81	orbit 0.53
rutger 0.86	0.00	public 0.98	doctor 0.72	man 0.41
church 0.66	0.00	govern 0.93	patient 0.68	cost 0.35
word 0.66	0.00	secur 0.90	diseas 0.61	launch 0.35
bibl 0.64	0.00	clipper 0.87	treatmnt 0.61	<b>system 0.35</b>
faith 0.64	0.00	chip 0.85	medicin 0.58	mission 0.32
christ 0.63	0.00	peopl 0.79	physician 0.50	flight 0.30
jesu 0.60	0.00	comput 0.69	food 0.50	henri 0.30

Table 1: Five ICs inferred from a document collection of 4 Newsgroup messages: four can be recognised as independent topics + a blank component separates out a common semantic noise (unsaid words).

- Removing the blank component has the effect of expanding the text with semantically related words.

'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 0.63 devic 0.62
'peopl' 'effect' 'diseas' 'medicin' 'diagnos'
medic 0.98 doctor 0.77 patient 0.75 treatment 0.71 physician 0.66 food 0.66 symptom 0.65 med 0.65 diet 0.65
'peopl' 'sin' 'love' 'christ' 'rutger' 'geneva' 'jesu'
god 0.99 christian 0.99 church 0.79 word 0.79 bibl 0.78 faith 0.78 agre 0.74 accept 0.73 scriptur 0.73
'peopl' 'public' 'system' 'agre' 'faith' 'accept' 'christ' 'teach' 'clinic' 'mission' 'religion' 'jesu' 'holi' 'doctrin' 'scriptur'
god 0.05 christian 0.05 rutger 0.04 word 0.03 church 0.03 bibl 0.03 love 0.03 man 0.03 truth 0.03
'govern' 'peopl' 'christ' 'food' 'rutger' 'church' 'atho'
god 0.74 christian 0.74 word 0.66 accept 0.64 bibl 0.64 faith 0.64 jesu 0.63 agre 0.63 effect 0.63

Table 2: Expansion of randomly picked documents from the 4 Newsgroups collection. The first line gives the list of words actually present in the document, followed by the top list of additional words that result by removing the blank component (along with their posterior probability).

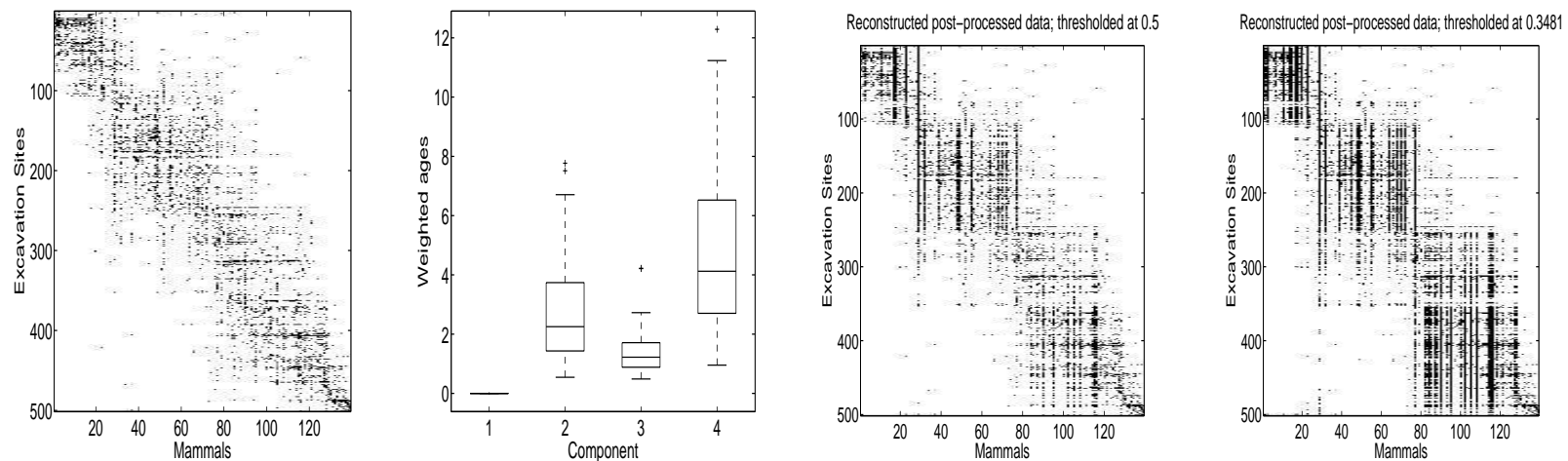
## Age discovery and missingness detection in paleontological data

- records of 139 mammal genera vs 501 sites of excavation
- 3 components capture contiguous disjoint time periods + 1 blank component indicating a different reason (factor) for non-records.
- Often, remains of a mammal are not observed at a site even though it probably lived there, as the preservation, recovery and identification of fossils are subject to random effects. According to palaeontologists<sup>a</sup>, an indication of missingness can be derived from the age order of the sites: if a mammal is observed at two sites but not at an intermediate site, it is possible (although not certain) that an observation at the intermediate site is missing.

---

<sup>a</sup>Professor Mikael Fortelius, University of Helsinki, personal communication.

- After removing the blank component, and thresholding at 0.5 (see Figure 6, third plot from the left), 1369 of the ‘likely to be missing’ records are filled in. Furthermore, by thresholding at an (estimated) value 0.3481, this number raises to 3642. The continuity of mammals as recovered by our binary ICA is now quite apparent on the rightmost plot of Figure 6.



**Figure 6:** From left to right: The palaeontological data, both the sites and the remains of mammals are ordered by age, for the ease of visual analysis of the results; Distributions of ages of mammals, weighted by  $\langle b_{kn} \rangle$ , for each component; Binary reconstruction of the absences in the data after having removed the noise component, using a threshold of 0.5 – these are superimposed with the observed presences; Binary reconstruction, when using an estimated threshold.

# Automatic finding of treatment groups from DNA fingerprints

- Initially a 4-class classification problem on binary data
- Binary ICA discovers the 4 classes in an unsupervised way
- No noise component is detected in this data

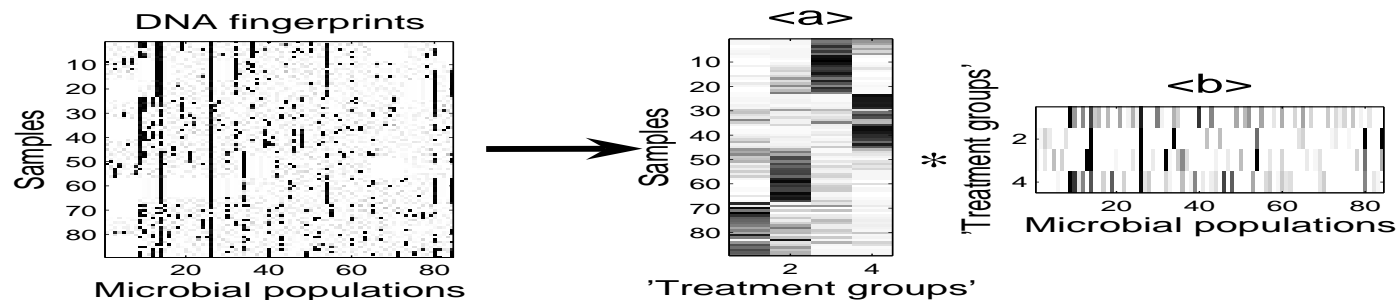


Figure 7: The estimated binary ICA mixing coefficients correlate with the treatment groups in DNA fingerprints.

## Conclusions

- ▶ We devised a variational ICA method for binary data
- ▶ Beta components are interpretable as grey-scale representations of the binary data
- ▶ Noise factors are separated out from systematic factors, and this allows us to detect and remove them
- ▶ Human input is taken into account in a principled and straightforward manner
- ▶ Applicable in various domains.