Pattern Anal Applic
DOI 10.1007/s10044-007-0096-4

THEORETICAL ADVANCES

The aspect Bernoulli model: multiple causes of presences and absences

Ella Bingham · Ata Kabán · Mikael Fortelius

Received: 18 April 2007 / Accepted: 14 October 2007

Abstract We present a probabilistic multiple cause model for the analysis of binary (0–1) data. A distinctive feature of the aspect Bernoulli (AB) model is its ability to automatically detect and distinguish between "true absences" and "false absences" (both of which are coded as 0 in the data), and similarly, between "true presences" and "false presences" (both of which are coded as 1). This is accomplished by specific additive noise components which explicitly account for such non-content bearing causes. The AB model is thus suitable for noise removal and data explanatory purposes, including omission/addition detection. An important application of AB that we demonstrate is data-driven reasoning about palaeontological recordings. Additionally, results on recovering corrupted handwritten digit images and expanding short text documents are also given, and comparisons to other methods are demonstrated and discussed.

Keywords Data mining · Probabilistic latent variable models · Multiple cause models · 0–1 data

1 Introduction

In multivariate binary data, only the presence (1) or absence (0) of each attribute is known, in contrast to count data where the actual frequencies of attribute occurrences are taken into account. Binary data arise in various applications, ranging from information retrieval, link analysis, transaction analysis and telecommunications to bioinformatics, to name a few. In this paper, we concentrate on probabilistic latent variable modelling of multivariate binary data, meaning that we aim at estimating the properties of the underlying system that has generated the observed data. It is assumed that the data arise due to latent (hidden) causes and their combinations. Revealing these causes gives new insight into the underlying system, and enables one to characterise the data in a compressed form. Probabilistic latent variable modelling is typically unsupervised, i.e. no "training data" with known latent causes are available.

Multiple cause models, termed also as factor models or distributed models ([1–5] and others) allow for several explanatory variables for each observation vector. That is, the elements of a vector-valued observation may have different underlying causes. In terms of clustering, an observation may belong to several clusters simultaneously.

We present a probabilistic, multiple-cause latent variable model for binary data. The aspect Bernoulli (AB) model, previously presented in a short preliminary version [6], can formally be seen as a Bernoulli analogue of the multinomial decomposition model known under the names

A part of the work of Ella Bingham was performed while visiting the School of Computer Science, University of Birmingham, UK.

E. Bingham (✉)
Helsinki Institute for Information Technology,
University of Helsinki and Helsinki University of Technology,
P.O. Box 68, 00014 Helsinki, Finland
e-mail: ella@iki.fi

A. Kabán
School of Computer Science, University of Birmingham,
Birmingham, UK
e-mail: a.kaban@cs.bham.ac.uk

M. Fortelius
Division of Palaeontology, University of Helsinki,
Helsinki, Finland
e-mail: mikael.fortelius@helsinki.fi

of aspect model, PLSA [3], and their generative versions such as latent Dirichlet allocation (LDA) [5] and multinomial principal component analysis (MPCA) [7]. Contrarily to multinomial models, where the event space is the set of attributes, for AB, the event space is the set [presence (0), absence (1)]. For a comprehensive exposition of event models for discrete data focusing on the difference between the independent Bernoulli and the multinomial event models in the context of text encoding see McCallum and Nigam [8]. A characteristic feature of the AB model is that noise in this event space is separated into one or a few distinct components, and this may further be straightforwardly exploited for noise removal.

Multiple-cause models for binary data have been devised before in the literature. Most notably, Saund's model [2] asserts an interaction model for the 1s in the data, which takes the form of a noisy-OR. However, the 0s are suppressed this way, and observing 0s remains a default uninteresting event. By contrast, in a linear Bernoulli model, the 0s and 1s are interchangeable. Keeping our model linear provides symmetry to AB enabling the analysis of the causes behind not only the ones (presences) but also the zeros (absences) in the data. Indeed in many applications it is of interest to model the zeros as well, when it comes to inferring hidden causes, as the absence (0) of an attribute might be indicative of an important underlying cause of interest. To give an illustrative example, the semantic content of two images that contain the digits '3' and '8', respectively, differ by pixels that are 'off' rather than 'on'. In various situations we may also encounter noise factors, which exclusively generate 0s, "wiping off" some of the content-bearing 1s. This is the case in text document data, where certain attributes (words) are genuinely absent, i.e. they have no intersection with the topical content of the observation (document) whereas others are absent for no specific reason other than the document is short. Similarly, black-and-white images may contain corrosion which turns a black pixel (1) into white (0). Stated briefly, there might be two kinds of zeros, which of course look the same in the data: "true absences" which agree with the content of an observation, or "false absences" (omitted presences), which might well have been 1s but due to some underlying cause remain unobserved. We have no prior knowledge about whether or not a data set under study contains such distorted observations and it is of interest to infer this from the data. As we will see, this is what the AB model is designed for. It enables us to automatically detect and distinguish between these two types of zeros under the AB model's generative assumptions. Detecting omitted presences may help, e.g. in query expansion in which short documents can be augmented by topically related words, or in image restoration by detecting the corrupted pixels. Clearly, by symmetry, the AB model can also distinguish between "true presences" (which are in accordance with the content of the observation) and "false presences" (which are due to a noise cause which explicitly turns 0s into 1s). This may be of use, e.g. in text-based forensic investigations.

In addition to the mentioned potential uses, in this paper, we demonstrate the abilities of the AB model in an actual application, in the context of palaeontological data [9] consisting of remains of mammal genera found at various sites of excavation across Europe and Asia. We may conjecture that there are underlying causes that can explain this data, such as those that reflect the communities of genera. Furthermore, if remains of a mammal genus were found on a site, we can infer that the mammal lived at or near that site. However, if no remains of a mammal genus were found, what can we infer? The palaeontological data are inherently noisy: It might be that remains of a genus are not recorded at a particular site even though the genus lived in the location of the site. As such, the data demands a model that is able to distinguish between true absences and false absences, both of which are coded as "0". We will show that the AB model is suitable for these purposes.

In addition to the actual palaeontological application, we will also demonstrate results on black-and-white raster images and binary coded text in order to assess the noise detection and removal performance on systematic and controlled experimental settings.

Our AB model can formally be seen as a special case of a more general matrix factorisation theory discussed, e.g. by Srebro [10]. It can also be seen as a special case of the models proposed for collaborative filtering by Hofmann and Puzicha [11] and Hofmann [12] and the URP model of Marlin and Zemel [13], if the observations are to be restricted to 0/1. A more complete review of related models will be given in Sect. 2.4. However, while these frameworks are formally closely related to our approach, our inferential scope is rather different. Our purpose here is to devise an appropriate model for reasoning about 0–1 data by detecting and separating out interpretable content-bearing factors, as well as "noise" factors in an automated manner. Separating out noise factors is quite important because when such factors are detected, they can subsequently be removed from the data. There is no readily available algorithm for this task, since most of the denoising literature is concerned with real-valued data. Secondly, it is also of interest here to study how such a specific instantiation of factorisation models compares to other models of 0–1 data, in terms of prediction and generalisation on real-world data.

Before proceeding, we make a note regarding the use of a number of almost synonymous terms in the paper: "aspect", "cause", "component", "factor", "prototype" and "basis". To avoid confusion, in this paper, we will

follow certain guidelines in the term usage. First of all, "cause" refers to a true underlying phenomenon in the data. In general, the causes are modelled by "components" which can further be characterised as follows. A component is called a "factor" in factor models, a group of models in which aspect models belong to, and hence the term "aspect" refers to a component of a linear convex factor model. A "prototype" is a component that has an interpretable representation, e.g. a cluster-centre. In turn, the term "basis" refers to the coefficients of the linear combination for a particular component, which may or may not be directly interpretable.

This paper is organised as follows. We first describe the model and place it in the context of various other models in Sect. 2. Experimental results are shown in Sect. 3, and Sect. 4 draws some conclusions and discusses possible future directions.

2 The model

In this section, we first describe the data generation process assumed in the AB model, and derive an implementation-friendly algorithm for estimating the model parameters. We then discuss related work and place AB in a broader context.

2.1 Derivation of the algorithm

We start by describing the data generation process of the aspect Bernoulli model. The indices $n = 1,\ldots,N$, $t = 1,\ldots,T$ and $k = 1,\ldots,K$ are used to index the observations, attributes and latent aspects, respectively. Let $\boldsymbol{x}_n$ denote a T-dimensional multivariate binary observation and x_{tn} the value of its tth attribute. The elements x_{tn} may be generated by different latent aspects k with probabilities specific to observation n and aspect k. The nth observation vector $\boldsymbol{x}_n$ is assumed to be generated as follows:

- Pick a discrete distribution $p(1|n),\ldots,p(K|n)$ over all the latent aspects $k = 1,\ldots,K$. The distribution is picked uniformly from the set of all such distributions.
- Separately for each element x_{tn} of $\boldsymbol{x}_n$, the following two steps are taken:
  - Pick a latent aspect k with probability $p(k|n)$
  - Let the latent aspect k generate 1 (presence) or 0 (absence) of the tth attribute. The Bernoulli probability of generating 1, $p(1|k,t)$, only depends on k and t and not on the observation index n.

Thus, there are two sets of unknown probability parameters in the model. Let us denote by $s_{kn} = p(k|n)$ the probability of choosing a latent aspect k in observation[1] n, and by $a_{tk} = p(1|t,k)$ the Bernoulli probability of the tth attribute being "on" conditioned on the latent aspect k. As K is typically significantly smaller than N, the total number $(T \cdot K + K \cdot N)$ of unknown parameters is smaller than the size $(T \cdot N)$ of the original data set, allowing for a compressed representation of the data.

In addition, a "dummy" indicator variable z_{tn} will denote which of the latent aspects generated the 0/1 event at the tth attribute of the nth instance: $\delta(z_{tn} - k)$ will equal one for exactly one aspect k, and $\delta(z_{tn} - k') = 0$ for all $k' \neq k$. We will use the shortcut $z_{tnk} = \delta(z_{tn} - k)$.

Summarising the generative process, we have the following dependency structure in the complete data likelihood of an instance n

$$p(\boldsymbol{x}_n, \boldsymbol{z}_n, \boldsymbol{s}_n|\boldsymbol{a}) = p(\boldsymbol{s}_n) \prod_{t=1}^{T} p(z_{tn}|\boldsymbol{s}_n) p(x_{tn}|z_{tn}, \boldsymbol{a}_t) \tag{1}$$

where $\boldsymbol{s}_n = (s_{1n},\ldots,s_{Kn})$ are the probabilities of selecting one of the K aspects, $\boldsymbol{a} = (a_{11},\ldots,a_{TK})$ are parameters of the model, consisting of the Bernoulli probabilities and $\boldsymbol{z}_n = (z_{1n1},\ldots,z_{Tn1},\ldots,z_{TnK})$. Further, we have

$$\begin{aligned} p(\boldsymbol{s}_n) &= \mathcal{U}_{\triangle}(\boldsymbol{s}_n) \\ p(z_{tn}|\boldsymbol{s}_n) &= \prod_k s_{kn}^{z_{tnk}} \\ p(x_{tn}|z_{tn}, \boldsymbol{a}_t) &= \prod_k \left[a_{tk}^{x_{tn}} (1 - a_{tk})^{1 - x_{tn}}\right]^{z_{tnk}} \end{aligned}$$

where $\mathcal{U}_{\triangle}$ is a uniform distribution on a simplex. The graphical representation as a plate diagram is shown in Fig. 1. The model assumes that the elements x_{tn} of $\boldsymbol{x}_n$ are conditionally independent given the latent variable $\boldsymbol{s}_n$. This is a standard assumption in generative modelling, and it signifies that all dependencies that exist in the observations are meant to be explained by the hidden variables of the model.

Thus, the complete data likelihood (1) now reads as

$$p(\boldsymbol{x}_n, \boldsymbol{z}_n, \boldsymbol{s}_n|\boldsymbol{a}) = \mathcal{U}_{\triangle}(\boldsymbol{s}_n) \prod_{t=1}^{T} \prod_{k=1}^{K} \left[s_{kn} a_{tk}^{x_{tn}} (1 - a_{tk})^{1 - x_{tn}}\right]^{z_{tnk}} \tag{2}$$

and so the probability of a data point under the model is obtained by marginalising the hidden variables:

$$p(\boldsymbol{x}_n|\boldsymbol{a}) = \int \mathrm{d}\boldsymbol{s}_n \mathcal{U}_{\triangle}(\boldsymbol{s}_n) \sum_{\boldsymbol{z}_n} \prod_{t=1}^{T} \prod_{k=1}^{K} \left[s_{kn} a_{tk}^{x_{tn}} (1 - a_{tk})^{1 - x_{tn}}\right]^{z_{tnk}} \tag{3}$$

$$= \int \mathrm{d}\boldsymbol{s}_n \mathcal{U}_{\triangle}(\boldsymbol{s}_n) \prod_{t=1}^{T} \sum_{k=1}^{K} s_{kn} a_{tk}^{x_{tn}} (1 - a_{tk})^{1 - x_{tn}} \tag{4}$$

[1] Note that at each attribute t of observation n, the latent aspect k is sampled *anew* from the distribution $p(1|n),\ldots,p(K|n)$.

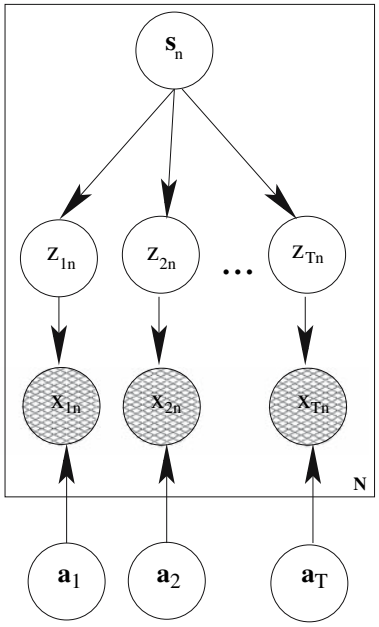

Fig. 1 Plate diagram representation of the aspect Bernoulli model. The textured nodes represent observed variables, the white nodes are unknown variables. Of these, those that have no parent nodes are treated as parameters of the model

where the summation in the first row is taken over all possible combinations of the z_{tnk}, and in the second row we have used the fact that only one of z_{tnk} equals 1 at each pair t and n.

This integral is analytically intractable for computing; therefore, the posterior distribution $p(\boldsymbol{s}_n|\boldsymbol{x}_n,\boldsymbol{a})$ is intractable as well (since its normalisation factor is exactly the above integral). A variety of approximate methods are available to use, such as the maximum a posteriori (MAP) point estimate, variational mean estimates or sampling-based methods.

It is outside the scope of this paper to analyse or compare the various possible estimation methods—for such a comparison in a fairly general setting of discrete latent variable models see Buntine and Jakulin [14]. For the purposes of this paper, we derive MAP estimates. This is the maximiser of the true posterior (the most probable $\boldsymbol{s}_n$ for each n) and it can be computed without requiring the availability of the full posterior of $\boldsymbol{s}$.

Of course, we should note that in general, one needs to be careful and aware that MAP estimates are prone to overfitting, especially when the data available for the estimation are scarce. However, as we demonstrate in the experimental section, AB being a factor-type model rather than a mixture over high-dimensional data, we did not encounter severe overfitting problems for a number of data sets analysed. In situations of excessively scarce data in turn, approximate Bayesian methods should be pursued to avoid overfitting. One could then treat the uniform density as a Dirichlet with all hyperparameters equal to 1, and employ the variational techniques developed in LDA [5], URP [15] or MPCA [7]. Indeed, we pursued some of these techniques for binary data analysis elsewhere [16, 17].

Since we have a uniform prior on $\boldsymbol{s}$, the maximum argument of the posterior coincides with the maximum argument of the likelihood; in other words, the MAP solution coincides with the maximum likelihood (ML) solution:

$$\begin{aligned}\boldsymbol{s}_n &= \operatorname*{argmax}_{\boldsymbol{s}} p(\boldsymbol{s}|\boldsymbol{x}_n,\boldsymbol{a}) = \operatorname*{argmax}_{\boldsymbol{s}} p(\boldsymbol{x}_n,\boldsymbol{s}|\boldsymbol{a}) \\ &= \operatorname*{argmax}_{\boldsymbol{s}} p(\boldsymbol{x}_n|\boldsymbol{s},\boldsymbol{a}) = \operatorname*{argmax}_{\boldsymbol{s}} \log p(\boldsymbol{x}_n|\boldsymbol{s},\boldsymbol{a})\end{aligned} \tag{5}$$

The same expression needs to be maximised also in $\boldsymbol{a}$, for parameter estimation. Expanding these expressions, we need to maximise the following, as a function of $\boldsymbol{a}$ and $\boldsymbol{s}_n$, $\forall$ $n = 1,\ldots,N$:

$$\sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{s}_n,\boldsymbol{a}) = \sum_{n=1}^{N}\sum_{t=1}^{T} \log \sum_{k=1}^{K} s_{kn} a_{tk}^{x_{tn}} (1-a_{tk})^{1-x_{tn}} \tag{6}$$

subject to the constraint $\sum_k s_{kn} = 1$ and $a_{tk} \in [0,1]$.

There is no closed-form solution and so we carry out this maximisation iteratively, making use of the EM methodology. Details are given for completeness in Appendix A.

The resulting EM algorithm is then the following: initialise all $\boldsymbol{s}_n$ and $\boldsymbol{a}$ within the required range. Then, iterate till convergence:

E-step:

$$q_{k,t,n,x_{tn}} = \frac{s_{kn} a_{tk}^{x_{tn}} (1-a_{tk})^{1-x_{tn}}}{\sum_\ell s_{\ell n} a_{t\ell}^{x_{tn}} (1-a_{t\ell})^{1-x_{tn}}} \tag{7}$$

M-step:

$$s_{kn} = \sum_t q_{k,t,n,x_{tn}}/T \tag{8}$$

$$a_{tk} = \frac{\sum_n x_{tn} q_{k,t,n,x_{tn}}}{\sum_n q_{k,t,n,x_{tn}}} \tag{9}$$

where $q_{k,t,n,x_{tn}} = p(z_{tn} = k|x_{tn},\boldsymbol{s}_n,\boldsymbol{a}_t)$.

2.2 Discussion and an implementation-friendly rewriting

Let us now analyse the above model in more detail. To start with, consider the likelihood of a single multivariate Bernoulli: $\prod_t p_t^{x_{tn}} (1-p_t)^{1-x_{tn}}$ where $p_t = p(x_{tn} = 1)$ is the probability for observing 1 in the tth element of any observation vector $\boldsymbol{x}_n$. A well-known extension of this is the single-cause mixtures of Bernoulli (MB) model

[18, 19] $\sum_k \pi_k \prod_t a_{tk}^{x_{tn}}(1-a_{tk})^{1-x_{tn}}$ where $a_{tk} = p(1|t,k)$ and π_k is the prior probability of the kth mixture component. Now let us instead extend the original simple parametric model in another vein, giving each observation vector n its own set of parameters $p_{tn} = p(x_{tn} = 1)$. This is clearly an over-parameterisation, so let us restrict it into a convex combination $p_{tn} = \sum_k a_{tk}s_{kn}$ where $\sum_k s_{kn} = 1$ and $0 \leq a_{tk} \leq 1$ for all t and k. This is indeed the core of the AB model, and we see this by rewriting:

$$p(\boldsymbol{x}_n|\boldsymbol{s}_n,\boldsymbol{a}) = \prod_{t=1}^{T}\sum_{k=1}^{K} s_{kn}a_{tk}^{x_{tn}}(1-a_{tk})^{1-x_{tn}} \tag{10}$$

$$= \prod_{t=1}^{T}\left(\sum_{k=1}^{K} a_{tk}s_{kn}\right)^{x_{tn}}\left(1-\sum_{k=1}^{K} a_{tk}s_{kn}\right)^{1-x_{tn}}. \tag{11}$$

To see the equality between (10) and (11), note that when $x_{tn} = 1$, then according to both (10) and (11) we have that $p(x_{tn}|\boldsymbol{s}_n) = \sum_k a_{tk}s_{kn}$; and for the case when $x_{tn} = 0$ we have $p(x_{tn}|\boldsymbol{s}_n) = \sum_k (1-a_{tk})s_{kn}$ from both (10) and (11). In obtaining the latter equality, we have used the convexity of the combination—note that $1 - \sum_k a_{tk}s_{kn} = \sum_k(1-a_{tk})s_{kn}$.

The likelihood in (11) indeed resembles the well-known Bernoulli likelihood if we denote by p_{tn}: $= p(x_{tn} = 1|\boldsymbol{s}_n) = \sum_k a_{tk}s_{kn}$ the Bernoulli probability of obtaining 1. Thus, the Bernoulli mean is factorised in a convex combination—which is a useful insight for relating this model to other distributed models of 0–1 data, as will be seen in the experimental section.

We can also rewrite the (7)–(9) to gain savings in the memory requirements and obtain a more convenient implementation, using the following observations.

1. The M-step updates only require sums over $q_{k,t,n,x_{tn}}$. Since the data are binary, we can re-write the E-step update expression (7), by separating the terms in which $x_{tn} = 1$ and those in which $x_{tn} = 0$, yielding

$$q_{k,t,n,x_{tn}} = s_{kn}\frac{x_{tn}}{\sum_\ell a_{t\ell}s_{\ell n}}a_{tk} + s_{kn}\frac{1-x_{tn}}{1-\sum_\ell a_{t\ell}s_{\ell n}}(1-a_{tk}) \tag{12}$$

2. From the theory of EM we know that each of the three update equations of the EM algorithm, (7)–(9), taken individually, is guaranteed not to decrease the objective that we maximise, i.e. (6). In addition, note that the M-step update of s_{kn} does not depend on any other parameters except through q and similarly, the same holds for each a_{tk} update. We can therefore choose to perform an E-step after each of the M-step updates and we are still guaranteed not to decrease the objective. Thus, in each iteration, we will perform the list of updates (7) and (8), and (7) and (9), or equivalently (12) and (8), and (12) and (9).

The reason why this is convenient is that we can then express the effect of one E-step and one of the M-step updates with a single equation simply by replacing the expression on the r.h.s. of the E-step update (12) into the M-step update. Doing this for both M-step updates, i.e. combining (12) and (8) and again (12) and (9), yields the following:

$$s_{kn} = s_{kn}\left\{\sum_t \frac{x_{tn}}{\sum_\ell a_{t\ell}s_{\ell n}}a_{tk} + \frac{1-x_{tn}}{1-\sum_\ell a_{t\ell}s_{\ell n}}(1-a_{tk})\right\}/T \tag{13}$$

$$a_{tk} = a_{tk}\sum_n \frac{x_{tn}}{\sum_\ell a_{t\ell}s_{\ell n}}s_{kn}/c_{tk} \tag{14}$$

where the denominator is

$$c_{tk} = a_{tk}\sum_n \frac{x_{tn}}{\sum_\ell a_{t\ell}s_{\ell n}}s_{kn} + (1-a_{tk})\sum_n \frac{(1-x_{tn})}{1-\sum_\ell a_{t\ell}s_{\ell n}}s_{kn}. \tag{15}$$

As we see, the result is a multiplicative form update for both s_{kn} and a_{tk}, and by construction, both of these latter updates are guaranteed not to decrease the maximisation objective. Therefore, alternating these two multiplicative form updates will necessarily converge to a local optimum of the likelihood.

It may also be interesting to note that the obtained algorithm can also be derived as an alternating optimisation (details given in Appendix B). However, although this simpler derivation yields the same multiplicative form fixed-point equations, it does not reveal the convergence guarantee. This guarantee comes from the EM interpretation.

The benefit of the so rewritten version of the algorithm is that we need not explicitly compute and store the posteriors $q_{k,t,n,x_{tn}}$ for estimating $\boldsymbol{s}$ and $\boldsymbol{a}$. Moreover, this version can easily be expressed in matrix form, which is more convenient to implement. This is justified by similar arguments as discussed above: choosing to combine the group of $q_{k,t,n,x_{tn}}$ updates (performed for all k,t,n using all parameter values fixed at their current values) with the group of s_{kn} updates (performed for all k,n using all q values fixed at their previously obtained current values); then again the group of $q_{k,t,n,x_{tn}}$ with that of the a_{tk} updates is still guaranteed not to decrease the likelihood, since every single constitutive update has this guarantee.

2.2.1 Algorithm

The summary of the obtained algorithm in a matrix-form notation is listed below.

- Initialise **A** and **S** within the appropriate domains.
- Iterate until convergence

$$\bar{\mathbf{A}} = 1 - \mathbf{A} \tag{16}$$

$$\mathbf{S} = \mathbf{S} \otimes \left\{ \mathbf{A}^{\mathrm{T}} [\mathbf{X} \oslash \mathbf{AS}] + \bar{\mathbf{A}}^{\mathrm{T}} [(\mathbf{1} - \mathbf{X}) \oslash \bar{\mathbf{A}}\mathbf{S}] \right\} \tag{17}$$

$$\mathbf{S} = \mathbf{S} \oslash \mathbf{R} \tag{18}$$

$$\mathbf{A} = \mathbf{A} \otimes \{ [\mathbf{X} \oslash \mathbf{AS}] \mathbf{S}^{\mathrm{T}} \} \tag{19}$$

$$\bar{\mathbf{A}} = \bar{\mathbf{A}} \otimes \{ [(\mathbf{1} - \mathbf{X}) \oslash \bar{\mathbf{A}}\mathbf{S}] \mathbf{S}^{\mathrm{T}} \} \tag{20}$$

$$\mathbf{A} = \mathbf{A} \oslash (\mathbf{A} + \bar{\mathbf{A}}) \tag{21}$$

where $\mathbf{R}$ denotes the matrix of normalisation factors of elements $R_{kn} = \sum_{\ell} s_{\ell n}$, $\forall\, k$, and $\otimes$ and $\oslash$ denote element-wise matrix product and division, respectively.

2.3 Scaling

The scaling per iteration of the ML estimation of an AB is $\mathcal{O}(NTK)$. This is less convenient as the $\mathcal{O}(\#(\text{nonzero})\text{K})$ scaling of multinomial aspect models, which scale linearly in the number of nonzero attribute occurrences in the data. However, this is the price we have to pay for having an independent Bernoulli likelihood model conditioned on $\boldsymbol{a}$ and $\boldsymbol{s}$. Independent Bernoulli component models, with very few exceptions [20], typically do not scale better than AB: The scaling per iteration of the Bernoulli mixtures is the same $\mathcal{O}(NTK)$. Logistic PCA [21], a recently introduced nonlinear distributed model for binary data, discussed in some detail later, scales as $\mathcal{O}(NTK^3)$ due to the matrix inversions that it requires.

2.4 Relation to other models

So far we have seen that AB is a probabilistic linear multiple cause model for 0–1 data that factorises the mean of the Bernoulli distribution into a convex combination of hidden causes, and explains both the 0s and the 1s in the data. Let us then contrast these properties to those of other models.

Starting from the factorisation in (11), we can draw parallels to a number of other multiple cause models in which a somewhat similar factorisation of the mean of the data distribution takes place. Perhaps, the most well-known probabilistic model for binary data is the single cause mixtures of Bernoulli (MB) model [18, 19], already mentioned in Sect. 2.1; however, as a single-cause model it assumes that all elements of the multivariate observation share the same latent cause. The Logistic PCA model [21] and the models of Tipping [22] and Collins et al. [23] decompose the so-called natural parameter θ of the Bernoulli distribution as $\theta_{tn} := \sum_k a_{tk}\, s_{kn}$, and the Bernoulli mean is then obtained using the logistic function $p_{tn} = 1/(1 + \mathrm{e}^{-\theta_{tn}})$. The nonlinear logistic function gives more flexibility as the parameters a and s need not be probabilities but can take any real values. For this reason, these models fit well to the data. However, a disadvantage of these models is the loss of interpretability of the parameters a and s. In contrast, the parameters of the linear decomposition in the AB model allow for insightful interpretations, as will be demonstrated later in this paper.

Apart from these Bernoulli-type models, the PLSA (probabilistic latent semantic analysis) [3], LDA (latent Dirichlet allocation) [5] and MPCA (multinomial PCA) [7] models for multinomial data have been quite popular over the past few years. Similarly to AB, these can be viewed as models that factorise the mean-parameter vector of a multinomial sampling model into a convex combination of 'latent causes'. The generative process of AB is almost identical to that of LDA, except that rather than a multinomial sampling, AB employs a conditionally independent Bernoulli sampling (conditioned on the parameters). Although from the technical point of view this may seem like a rather small difference, it dramatically affects the type of data that AB is suited to analyse and thus the sort of inferences that it is meant to accomplish. In a multinomial sampling model over some attribute space, at each draw, the attribute that gets drawn is present, all others are necessarily absent. In turn, in our independent Bernoulli sampling model, given $\mathbf{a}$ and $\mathbf{s}$, the presence or absence information is sampled independently for each attribute. In other words, conditioned on the model parameters, the presence of an attribute does not tell us anything about the presence of another attribute, and several attributes can be present in the same time. Therefore, despite the formal similarity between AB and the PLSA, LDA or MPCA models, AB needs to model and 'explain' both the presence and the absence events for each attribute and each datum instance. Most of this paper is concerned with demonstrating what sort of useful information we can learn from binary data by doing such an analysis.

As already mentioned, AB could formally be seen as a special case of the URP model [15]. The URP model was designed for collaborative filtering, and it posits several conditionally independent multinomials (to model some discrete set of ratings), one for each product. Thus, with ratings restricted to 0/1, URP would reduce to AB—however, such a model has not been previously investigated. Previous related collaborative filtering methods have also been studied by Hofmann and Puzicha [24] and Hofmann [12]; the model presented in the latter can be used for arbitrary response scales.

Saund's model [2] is one of the first multiple cause models for binary data. It does not perform a linear decomposition of the Bernoulli mean parameter but instead

it identifies a nonlinear "noisy-OR" relationship between the hidden causes. A closed-form solution is not available but a gradient algorithm maximising the likelihood is given [2] and a mean-field approximate solution has been provided later [25]. A somewhat similar model is the topic model presented by Seppänen et al. [26]; there the relationship between latent causes is described by a discrete logical OR function. The problem of finding an optimal topic assignment is shown to be NP-hard, and approximative iterative algorithms for the estimation of the parameters are given [26]. A discrete logical OR function is also discussed by Jaakkola [27] who gives upper and lower bounds for the likelihood.

Early approaches to multiple cause models have been presented by Barlow et al. [28], Földiák [29], Schmidhuber [30] and Zemel [31]; in these models the data are not necessarily assumed binary valued. Later, Dayan and Zemel [25] have presented a model where the latent components compete with each other and thus ensure that they account for representing different parts of the binary data space. Yet another formulation is given by Marlin and Zemel [13] in their multiple multiplicative factor models, also allowing different components to specialise to a subset of the data space; their models are given for multinomial data but can be easily adapted for binary. Recently, an interesting approach of latent class modelling in relational binary data has been presented by Kemp et al. [32].

Non-probabilistic methods for the analysis of binary data include the method of frequent sets [33] which as such does not give a model of the data but instead reveals local patterns of co-occurrence of attributes. Subspace clustering, also known as co-clustering or double clustering [34], analyses the structure of binary data and partitions the data both on the level of observations and on the level of attributes; in contrast to latent variable methods, no underlying causes are assumed to have generated the data, and no overlap between the clusters are allowed. Yet another method of unsupervised learning from 0–1 data is the famous Boltzmann machine [35].

Apart from binary data, well-known methods for factoring continuous data include principal component analysis (PCA) [36], independent component analysis (ICA) [4] and nonnegative matrix factorisation (NMF) [37–39]. Of these, NMF is perhaps the closest to our approach, as its decomposition reads $x_{tn} = \sum_k a_{tk} s_{kn}$ where a_{tk} and s_{kn} are nonnegative but not restricted to be probabilities. A probabilistic version of PCA is given by Tipping and Bishop [40] and further discussed, e.g. by Dahyot et al. [41].

Srebro and Jaakkola [42, 10] and Gordon [43] discuss the general class of matrix factorisations and give an overall view to the problem. Haft et al [44] present a latent variable model with binary sources and continuous data.

Having placed our approach in the more general context of matrix factorisation and multiple cause modelling literature, we now turn to further analyse and experimentally demonstrate the use of AB on real world data sets, contrasting it to some of the related models reviewed here. In particular, AB turns out to be well suited to modelling high-dimensional 0–1 data and noise removal from 0–1 data. We will also analyse the representational tendencies of AB and other models through a number of examples and this contributes to better understanding of different matrix factorisation models in general and the AB model in particular.

3 Experiments

In this section, we first describe the data sets used in the experiments. Model selection in terms of choosing an optimal number of latent aspects is then addressed, followed by detailed analyses of the representation tendencies of the AB model, the interpretability of model parameters and model's ability to detect and remove discrete 0–1 noise, such as 'omissions' and 'additions' of presences or absences.

In the experiments, the AB model is compared to mixtures of Bernoulli (MB) [18, 19], probabilistic latent semantic analysis (PLSA) [3], logistic PCA (LPCA) [21] and nonnegative matrix factorisation (NMF) [39], when appropriate. Of these, the first two are estimated using an EM algorithm; the LPCA model is estimated by alternating least-squares optimisation; and NMF is estimated by a multiplicative update scheme that optimises an Euclidean distance. The methods are implemented in Matlab, in the form presented in the respective references.

3.1 Data sets

The data sets used to demonstrate the performance of the aspect Bernoulli model are quite distinct in their nature: palaeontological findings of mammals at various sites of excavation; black-and-white images of handwritten digits; and binary coded newsgroup documents.

3.1.1 Palaeontological data

Our palaeontological data come from the NOW database, a public resource based on a collaboration between mammal palaeontologists.[2] The NOW data derive from the published

[2] NOW: Neogene of the Old World, http://www.helsinki.fi/science/now/.

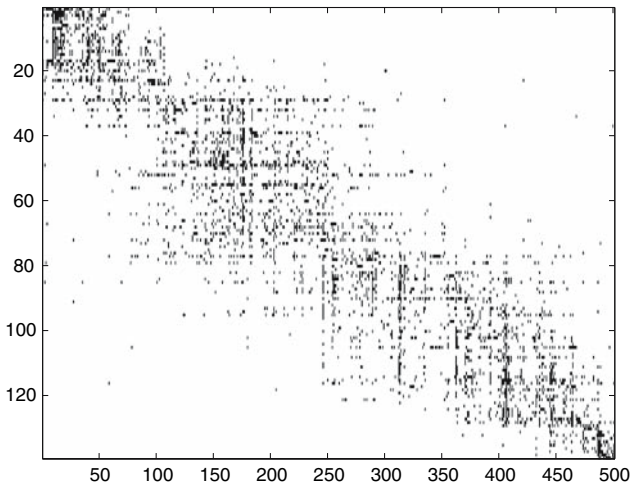

Fig. 2 Palaeontological data: rows correspond to genera and columns to sites of excavation

literature as well as from unpublished compilations by contributors.

The data set we use comes from NOW public release 030717. We have excluded small mammals (orders Insectivora, Chiroptera, Lagomorpha and Rodentia), and limited the geographic coverage to Europe, arbitrarily truncated towards Asia at 60° eastern longitude. Our data set consists of 501 sites (localities where fossils have been recovered, usually by excavation), in which occurrences of 139 genera are observed. Genera with less than 10 occurrences and sites with single genera have been excluded. We interpret the fossil sites as observations and the genera as attributes. The data are quite sparse: 5.08% of the entries are 1. The data matrix is seen in Fig. 2.

In addition, we have access to the ages of the fossil sites. The age is estimated from all available evidence, including, at best, radiometric dating and palaeomagnetism, but the majority of the sites are dated by means of mammal biochronology, i.e., the evolutionary change observed in the mammals themselves. For technical details of how age is handled in NOW see Fortelius et al. [9] or the NOW web site. The age estimates in our data set vary between 2 and 23 million years. The age information will be used to validate and visualise the results shown later.

The palaeontological data are inherently noisy: it might be that remains of a genus are not recorded at a particular site even though the genus lived in the location of the site. There are a number of reasons why an observation may not be recorded in the data. Sampling plays a major role: in small samples, only the most common genera tend to be recorded, and the number of rarer genera present continues to increase with sampling for most represented sample sizes [9]. The preservation, recovery, and identification of fossils are all random to some extent; in addition, there are more systematic reasons for spurious absences. Mammals differ in size and anatomy, and as a result some are more likely than others to be preserved and correctly identified. Sometimes, only one group of genera (e.g. the primates, the pigs) has yet been studied from a site. Similarly, the discovery of remains of common genera is rarely published without some particular reason, such as new discoveries of more rare ones. A third systematic reason is that a rare genus might not be recognised because no specialist was available. All these phenomena incur absences of attributes in the data.

The NOW data used here are quite typical of palaeontological data sets; if anything, most data sets are even more sparse. From a palaeontologist's point of view, the possibility to distinguish between "true absences" and "false absences" therefore has great appeal, along with other methods that strive to compensate for the low level of sampling (e.g. [45, 46]). Our AB analysis may provide new insights into this issue, as will be demonstrated in the following.

3.1.2 *Black-and-white raster images*

Another data set considered for studying the performance of the aspect Bernoulli model is a collection of 2000 binary digit images of handwritten numerals.[3] There are 200 instances from each digit category ('0', '1',…,'9'), each image containing 15 × 16 pixels, each of which can be either "on" (1) or "off" (0). In the original setting, any pixel that is off can be explained by the content of the image and is thus a "true absence". We later add corrosion-like new causes to the observed pixel values in the data, by randomly turning some pixels off or on. This data set is thus suitable as a basis for controlled experimental validation. Especially, we will demonstrate the performance of AB and several other methods in correcting for such corrosion.

3.1.3 *Binary coded text*

The third real-world data set is a subset of the 20 newsgroup corpus:[4] short Usenet messages from four newsgroups 'sci.crypt', 'sci.med', 'sci.space' and 'soc.religion.christian'. We selected 100 consecutive documents from each newsgroup and converted them into a binary term by document matrix using the Bow toolkit.[5] Text document data inherently contains omitted presences of words—not all

[3] http://www.ics.uci.edu/∼mlearn/MLSummary.html.

[4] http://www.cs.cmu.edu/∼textlearning/.

[5] http://www.cs.cmu.edu/∼mccallum/bow/.

words that may express a topic are covered in a document about that topic. Some documents are really short, made up by just a few words, and some longer ones utilise a richer dictionary. Typically, there is a dispersion of the richness from very concise to quite extensive documents in a collection, and of course, not the same words are omitted each time when expressing a given topic. Thus, obviously there may be different reasons why words do not appear—as well as there may be different reasons why they do. Revealing such ambiguities can be useful in, e.g. query based search. We note that previous statistical text modelling approaches have only been concerned with ambiguities created by presences of terms (not their absences!), such as synonymy and polysemy.

3.2 Model order selection

Here we consider the issue of how many components is the optimal choice. A number of model selection criteria are available to use. However, the optimal model order may depend on the application [47] and this is often overlooked in the machine learning literature. Of foremost importance in nearly all cases is the out-of-sample performance. Smyth [48] emphasises the use of cross validation for this reason. Generally, a model selection that reflects the objective of the modelling process should be adopted. For prediction problems, the model selection criterion should be based on the quality of predictions, whereas in data-explanatory tasks the aim is often related to Occam's philosophical principle, namely to finding the most parsimonious model that explains the data, but not simpler than that. The choice between prediction and explanation as the purpose for model selection is also discussed by Heckerman and Chickering [49] in the Bayesian model selection framework. We will consider two methods to cover both of these considerations within our frequentist approach.

3.2.1 *Cross-validation-based model selection for data prediction*

Let us first consider a model selection criterion for predictive purposes. Figure 3 shows the ten-fold cross-validated out-of-sample log likelihoods of the models investigated here, for all data sets. The out of sample likelihood is a measure that reflects the predictive capabilities of the models on these data. The procedure we are using is known as "empirical Bayes" and so we compute the empirical Bayes test likelihood, which is the following:

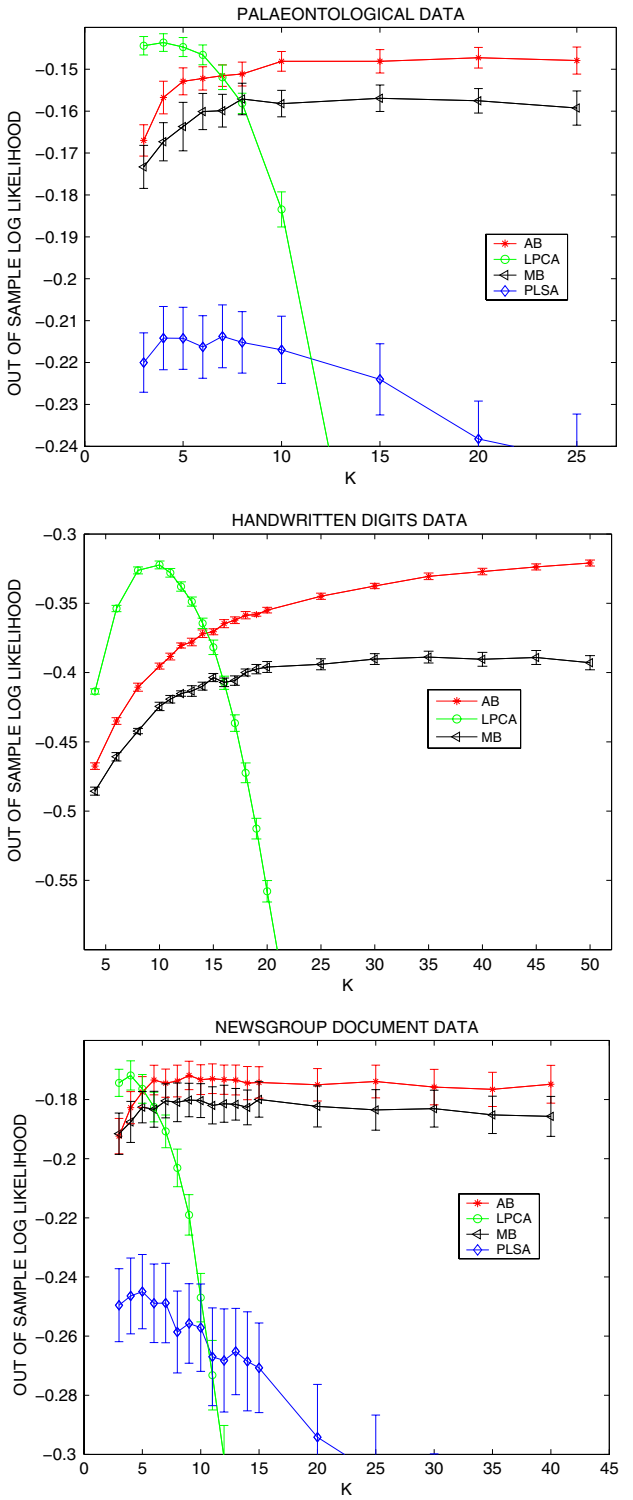

Fig. 3 Out-of-sample log likelihood for AB, LPCA, MB and PLSA, measuring the predictive capabilities of the models on the palaeontological data (*top panel*), handwritten digit data (*middle panel*) and newsgroup document data (*bottom panel*). *Horizontal axis* Model order (number of estimated components K). *Error bars* show one standard error on both sides of the mean of the folds in 10-fold cross-validation. In the *middle panel*, PLSA is below -3 and thus not shown

$$\log \int_{\boldsymbol{s}} p(\boldsymbol{x}_{\text{test}}|\boldsymbol{s},\boldsymbol{a})p(\boldsymbol{s})\mathrm{d}\boldsymbol{s} \tag{22}$$

where $p(\boldsymbol{s}) \approx 1/N_{\text{train}} \sum_n \delta(\boldsymbol{s} - \boldsymbol{s}_n)$ is the empirical sample density of the estimates of $\boldsymbol{s}$ as obtained from the training data of size N_{train} [50, 5]. For AB, the empirical Bayes test log likelihood associated with one test point is computed as the following:

$$\log \frac{1}{N_{\text{train}}} \sum_n \prod_t \left(\sum_k a_{tk} s_{kn} \right)^{x_{t,\text{test}}} \left(1 - \sum_k a_{tk} s_{kn} \right)^{1-x_{t,\text{test}}}, \tag{23}$$

for PLSA it reads

$$\log \frac{1}{N_{\text{train}}} \sum_n \prod_t \left(\sum_k a_{tk} s_{kn} \right)^{x_{t,\text{test}}} \tag{24}$$

and for LPCA, respectively

$$\log \frac{1}{N_{\text{train}}} \sum_n \prod_t \left(\frac{1}{1 + \exp\left(-\sum_k a_{tk} s_{kn}\right)} \right)^{x_{t,\text{test}}} \times \left(1 - \frac{1}{1 + \exp\left(-\sum_k a_{tk} s_{kn}\right)} \right)^{1-x_{t,\text{test}}}. \tag{25}$$

For MB, the test log likelihood does not involve a density over $\boldsymbol{s}$:

$$\log \sum_k \pi_k \prod_t a_{tk}^{x_{t,\text{test}}} (1 - a_{tk})^{1-x_{t,\text{test}}}. \tag{26}$$

In the above formulae, n ranges over the training points; specifically, s_{kn} are obtained for the training point $\boldsymbol{x}_n$. Instead, $x_{t,\text{test}}$ is the tth dimension of a new, previously unseen test point. The empirical test likelihood for an out of sample set of test points is then simply the average of the test likelihoods obtained for the individual test points.

Figure 3 shows the 10-fold cross-validated test likelihoods over a range of model orders. From these results, it is clear that AB consistently and significantly outperforms MB, except for the newsgroup data, in which the AB likelihood is higher but the error bars overlap. PLSA remains the poorest in this comparison, partially because its likelihood is computed differently: the zero entries of data do not contribute to the log likelihood (24) as $\log (\sum_k a_{tk} s_{kn})^x = 0$ when $x = 0$. One might expect NMF to behave similarly to PLSA (see Buntine [7] for a discussion of the similarity of NMF and PLSA)—in general, comparing the likelihoods of multinomial and Bernoulli models is problematic.

Interestingly, AB does not over-fit on these data sets over a wide range of model orders considered. (In the palaeontological data, over-fitting has been experienced after 30 components only.) In comparison with LPCA, AB requires more components but it achieves comparable performance. Given the cubic scaling of LPCA versus the multi-linear scaling of AB, as discussed in Sect. 2.3, AB may then be a preferable choice for modelling and analysis of binary data matrices. In addition, the most important advantage of AB is its intuitive data explanatory capability, which will be demonstrated in the next few sections. This is a consequence of the constrained nature of the AB parameters, which are all positive and probabilistic quantities, and thus easy to interpret. In turn, the LPCA parameters are unconstrained, resulting in a greater compression capacity but lack of interpretability.

3.2.2 AIC-based model selection for data explanation

Contrarily to prediction tasks, one often prefers a parsimonious data explanatory model. Following the arguments given by Ripley [47], a procedure designed to achieve this objective, in models estimated by maximum likelihood, is the Akaike Information Criterion (AIC) [51]:

$$\text{AIC}(K) = -\mathcal{L}(K) + P(K) \tag{27}$$

where K is the number of latent components, $\mathcal{L}$ is the in-sample log likelihood of the model and P is the number of free parameters that need to be estimated. In the AB model, $P(K) = TK + (K - 1)N$. The optimal model order is then found by minimising (27) under K.

For the palaeontological data, the AIC suggests $K = 4$ components; for the newsgroup data $K = 5$ and for the handwritten digit data $K = 15$.

3.3 Omission/addition detection by "phantom" latent aspects: an analysis

Here we provide some insights into the representational properties of the AB decompositions. In particular, we discuss the ability of the AB model to detect and distinguish between two types of zeros ("true absences" and "false absences") and similarly between two types of ones ("true presences" and "false presences"). These abilities were not discussed above when the model and algorithm were presented, and indeed the abilities are not "hard-coded" into the model. Instead, the detection of values that disagree with the topical content of an observation, namely false absences or presences, is accomplished by factors that we will call "phantom" latent aspects. A "white phantom" is a latent aspect which has a negligible probability of generating a value 1 at any attribute, meaning $a_{tk} \approx 0$ at all t, and thus it explicitly generates zeros in the data and can be used to detect and distinguish false absences from true absences. In contrast, a "black phantom" generates the

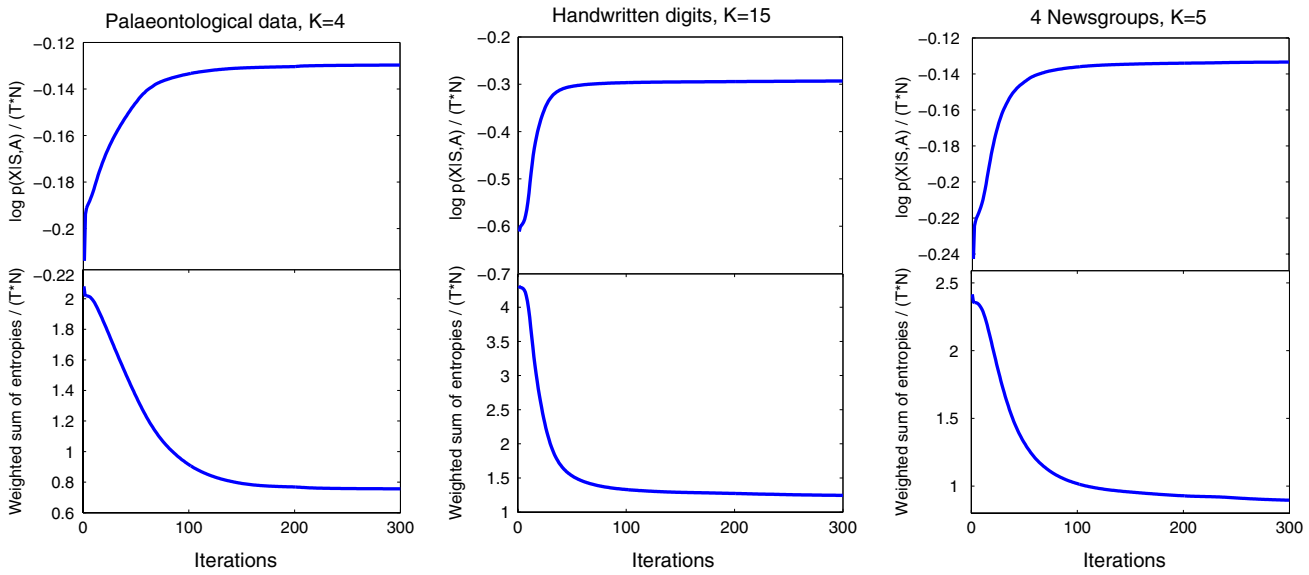

Fig. 4 *Top* Log likelihood of data (Formula 6). *Bottom* The weighted sum of parameter entropies (Formula 31). *Horizontal axis* EM iterations

value 1 at all attributes, $a_{tk} \approx 1$ at all t, and it can be used to distinguish added presences from true ones. We would like to stress that the phantoms are never imposed but instead found in the learning procedure when appropriate.

To provide an insight into this representation scheme, we analyse the implications of the optimisation performed by the EM algorithm, in terms of the entropies of the parameters involved.

At the stationary point of the log likelihood, using formulae (8)–(9), the sum of conditional expectations of the joint likelihood of the data and latent variables $\boldsymbol{z}_n$, conditioned on $\boldsymbol{s}_n$ and $\boldsymbol{a}$—i.e. the first part of the F-term in Eq. (38) in Appendix A—is the following.

$$\sum_n \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \log p(\boldsymbol{x}_n, \boldsymbol{z}_n | \boldsymbol{s}_n, \boldsymbol{a}) \tag{28}$$

$$= \sum_{n,k} \log s_{kn} \sum_t q_{k,n,t,x_{tn}} + \sum_{k,t} \log a_{tk} \sum_n x_{tn} q_{k,n,t,x_{tn}} + \sum_{k,t} \log(1 - a_{tk}) \sum_n (1 - x_{tn}) q_{k,n,t,x_{tn}} \tag{29}$$

$$= T \sum_{n,k} s_{kn} \log s_{kn} + \sum_{k,t} a_{tk} \log a_{tk} \sum_n q_{k,n,t,x_{tn}} + \sum_{k,t} (1 - a_{tk}) \log(1 - a_{tk}) \sum_n q_{k,n,t,x_{tn}} \tag{30}$$

$$= -T \sum_n H(\boldsymbol{s}_n) - \sum_{k,t} H([a_{tk}, 1 - a_{tk}]) \sum_n q_{k,n,t,x_{tn}} \tag{31}$$

which is a weighted sum of the entropies of the model parameters. Here we have used $q_n(\cdot) \equiv p(\cdot | \boldsymbol{x}_n, \boldsymbol{s}_n, \boldsymbol{a})$ and $p(\boldsymbol{x}_n, \boldsymbol{z}_n | \boldsymbol{s}_n, \boldsymbol{a}) = \prod_t \prod_k [s_{kn} a_{tk}^{x_{tn}} (1 - a_{tk})^{1 - x_{tn}}]^{z_{tnk}}$ and $\sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) z_{tnk} = q_{k,t,n,x_{tn}}$. The latter equality is obtained as detailed in Appendix A.

It is of interest now to follow how this weighted sum of entropies modifies during the EM iterations (7–9). Figure 4 shows the monitoring of (31) and the data likelihood (6) against iterations, for the data sets analysed.

We observe that the weighted sum of entropies decreases monotonically with the EM iterations. The decrease is very similar to the increase of the data log likelihood. This behaviour intuitively explains two representational tendencies of the model:

- a tendency towards a sparse distribution of $\boldsymbol{s}_n$ (only a few latent aspects are active at a time), due to the first sum of terms in (31)
- a tendency towards extreme binary values in a_{tk}, due to the second sum of terms in (31)

Specifically, in the extreme case when the data support that only one latent cause is active at a time, the representation reduces to a single-cause mixture; this implies that the bases a_{tk} are local averages of data. Averaging black and white (which is the case when a varying degree of omissions or additions are present in the data at random locations) would result in grey values in a_{tk}, i.e. high entropy Bernoullis—this is not preferable in the light of (31), so the method chooses to keep two active causes, namely one content-bearing aspect and one "phantom" aspect. The reduction of grey values in a_{tk} this way obtained compensates for the slight increase in the entropy of $\boldsymbol{s}_n$ when more than one s_{kn} become active for a given n.

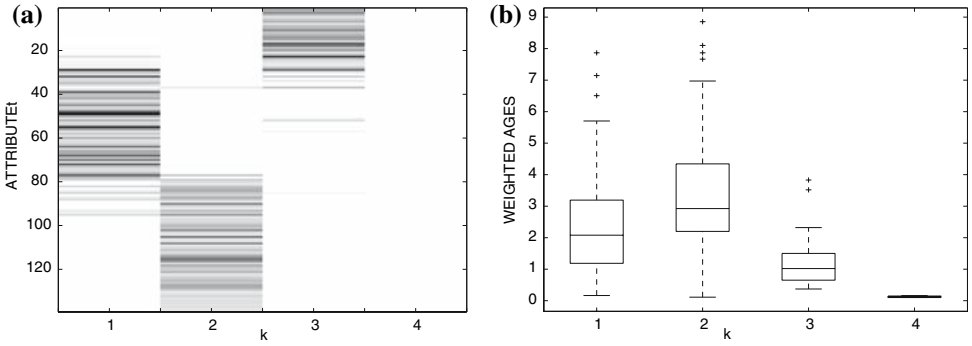

Fig. 5 **a** Values of the parameters $a_{tk} = p(1|k,t)$ given by the AB model at latent aspects $k = 1,\ldots,4$ and attributes (genera, sorted by their ages) $t = 1,\ldots,39$. **b** Distributions of weighted ages of genera in different latent aspects $k = 1,\ldots,4$. The age of genus t is weighted by the probability a_{tk}

A somewhat similar analysis has been useful for understanding the behaviour of other discrete variable models too [52–54].

Interestingly, an analogous derivation can be performed in the single-cause Bernoulli mixture model: the corresponding lower bound of the complete data likelihood can be written as $Q = -NH(\pi) - \sum_{k,t} H([a_{tk}, 1 - a_{tk}]) \sum_n s_{kn}$ where $\pi = (\pi_1,\ldots,\pi_K)$ is a vector of the prior probabilities of the mixture components and s_{kn} is the posterior probability of component k causing observation n. However, phantom-type components cannot arise as only one mixture component is allowed per observation and a phantom alone cannot explain both the ones and the zeros in the observation.

3.4 Parameter interpretability in palaeontological data

In this section, we will demonstrate that the modelling assumptions of AB give rise to quite intuitive and interpretable representations.

On the basis of the Akaike Information Criterion, the model order of $K = 4$ latent aspects was chosen for the analysis of palaeontological data. We now estimate the corresponding AB model. Figure 5a shows the values of the parameters a_{tk} giving the probability that the latent aspect k generates a value 1 at attribute (genus) t. White corresponds to zero probability and black to one. We can see that the aspects concentrate on distinct time intervals (the attributes in the data set are roughly ordered based on their ages). Also, there is one blank aspect to explain unknown false absences, giving a zero probability for all attributes. We call this kind of aspect a "white phantom". It generates zeros in the data, in contrast to other latent aspects that generate both zeros and ones.

To avoid ending up in a local minimum of likelihood, we estimate the model repeatedly with random initialisations and choose the in-sample likelihood-optimal values for presentation. A phantom such as the one in Fig. 5 was found in 28 out of 30 randomly initialised restarts.

Let us visualise the grouping of genera by drawing a box plot of the ages of genera captured by different latent aspects. Figure 5b shows for each latent aspect k the distribution of the ages of genera t weighted by the probabilities a_{tk}. We can see that different latent aspects indeed concentrate on different periods in time. The Wilcoxon rank sum test applied on all pairs of distributions indicates that they are distinct: the P values range between 0.0000 and 0.0201 for the null hypothesis of median equality.

The latent aspects can be viewed from a different angle if we consider the distributions s_{kn} giving the probability of latent aspect k being present in observation (site) n. The distributions are shown in Fig. 6a. The fourth aspect, the "phantom", is again different in its behaviour: it seems to have a nonzero probability in most observations. Thus, the model proposes that a phantom cause is present in a number

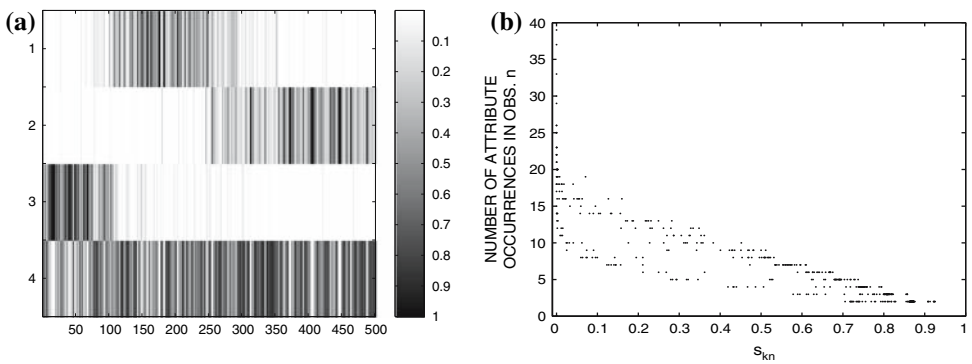

Fig. 6 **a** Distributions $s_{kn} = p(k|n)$ given by the AB model at latent aspects $k = 1,\ldots,4$ and observations (sites of excavation, sorted by their ages) $n = 1,\ldots,501$. **b** The value of s_{kn} versus the number of attribute occurrences in observation n for the phantom aspect k

of observations; by its presence, it generates absences of attributes, as seen in Fig. 5a. The varying probability of the phantom has a negative correlation with the number of ones per observation: The observations having only a few attribute occurrences have a large probability of the phantom being present, as seen in Fig. 6b.

The parameters a_{tk} and s_{kn} given by the LPCA, MB, PLSA and NMF models in turn (not shown) do not separate any blank cause to explain unknown false absences. Instead, the parameters given by MB, PLSA and NMF merely group with respect to time, quite similarly to the non-phantom parameters of the AB. The parameters given by LPCA, as well as the bias term included in the model, range across positive and negative values as they are not restricted to be probabilities but instead give the decomposition of the natural parameter θ of the Bernoulli distribution, up to rotation; the parameters are thus difficult to interpret.

3.5 Text document representation

We now turn to the newsgroup document data and demonstrate the latent aspects found by the AB model. The latent aspects can be visualised by listing for each aspect k the terms t having the largest probability a_{tk} of being generated by the aspect. We estimate $K = 5$ aspects suggested by the Akaike Information Criterion (27). Table 1 lists the keywords and their probabilities in descending order. The second aspect is a "phantom" aspect which gives a zero probability for the presence of any term. The other four are clearly related to the various topics of discussion. The model was estimated repeatedly, with random initialisations, and a phantom was found in 28 out of 30 initialisations.

The probabilities a_{tk} and s_{kn} in the newsgroup data behave quite similarly to those in the palaeontological data: for each aspect k, a group of terms t has a large probability a_{tk} of being "on", except for the phantom aspect. Respectively, each aspect k is active mainly in a subset of documents n, represented by the distributions s_{kn}, except for the phantom aspect which is active in most documents. This is seen in Table 1: the figures on top of each column k give $\sum_n s_{kn}$, the sum of the probabilities of the kth aspect in all documents; we see that the "phantom" aspect has a large overall probability compared to the other aspects.

In Table 1 we also note that in addition to the ambiguities regarding absences of terms, solved in the AB model in an original manner with the aid of "phantom" aspects, AB is also able to capture the well-known ambiguities that are associated with presences of terms—synonymy and polysemy. An example of synonymy can be noted in the given example within the *medical* aspect, where both 'medic' and 'doctor' are terms whose presence is highly probable. Polysemy is captured by that the presence of the same word may be generated by several topical aspects, e.g. the presence of the word 'system' is generated by both the *space-related* and *cryptographic* aspects. The aspect identifiers, shown in the table header, have intentionally been chosen as adjectives, in order to emphasise that the keyword lists represent in fact common features extracted from the corpus and are in general not cluster-centres. Naturally, if the corpus consists of well separated clusters then the main features will consequently be close to the cluster-centres, due to the clustering tendency of the model. However, the clustered structure is not artificially imposed, as in the case of single-cause mixtures. Indeed, e.g. the omission of words is a common feature of all text-based documents and this has been accounted for by the phantom topic.

Table 1 Five aspects k in a document collection of Usenet newsgroups sci.crypt, sci.med, sci.space and soc.religion.christian, presented as lists of terms t having the largest probabilities a_{tk} (shown after the terms)

Religious 45.1	*Phantom* 152.9	*Cryptographic* 42.9	*Medical* 48.2	*Space-related* 59.0
god 1.00	agre 1e-03	kei 1.00	effect 0.84	space 0.76
christian 1.00	sternlight 1e-11	encrypt 1.00	peopl 0.72	nasa 0.59
peopl 0.95	bless 3e-12	**system** 1.00	medic 0.66	orbit 0.49
rutger 0.81	truth 3e-15	govern 0.90	doctor 0.52	man 0.37
word 0.63	peopl 2e-15	public 0.89	patient 0.47	cost 0.35
church 0.63	comput 3e-16	clipper 0.84	diseas 0.42	**system** 0.34
bibl 0.61	system 9e-19	chip 0.83	treatment 0.40	pat 0.33
faith 0.60	man 1e-19	secur 0.82	medicin 0.40	launch 0.32
christ 0.59	nsa 1e-21	peopl 0.70	food 0.35	mission 0.30
jesu 0.56	shuttl 4e-22	comput 0.65	med 0.33	flight 0.28

Besides four aspects representing the topical features of discussion, there is an additional "phantom" aspect common to all documents, explaining absences of words which are not due to real topical causes. The top row gives $\sum_n s_{kn}$ reflecting the overall probability of aspect k

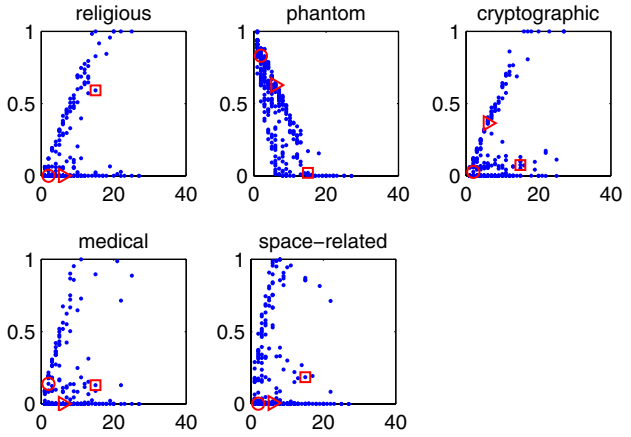

Fig. 7 The probability s_{kn} of the individual aspects (topics) plotted against the number of different words in document n (*horizontal axis*). The probability of the phantom-topic (aspect no. 2) is negatively correlated with the richness of the documents, whereas the real-topics are positively correlated with the richness. Three documents are highlighted: *circle* "system" "medicin"—a very short document; *square* "peopl" "public" "system" "agre" "faith" "accept" "christ" "teach" "clinic" "mission" "religion" "jesu" "holi" "doctrin" "scriptur"—a fairly long document with rich heterogeneous topical content; and *arrowhead* "govern" "secur" "access" "scheme" "system" "devic"—a medium size document focused on a single topic

Figure 7 depicts scatter plots of the probabilities s_{kn} of each aspect k against the number of distinct words which appear in the documents n, one subplot for each k. The probability of the phantom correlates negatively with the richness of the document. All real topical aspects in turn correlate positively with the richness of the documents. Also, as an example, three documents are highlighted. It is seen that despite s_{kn} sums to one w.r.t. k at each n, it is still possible to represent multiple causes of the same document n by letting s_{kn} take values in the whole range [0,1]. In contrast, in a single-cause model, we would have $s_{k'n} = 1$ for one $k = k'$ and 0 for other $k \neq k'$.

The analysis of individual documents is continued in Table 2. The first column lists the words t which are present in the document n, and in the second column the most probable aspects k for each word are given along with their posterior probabilities $q_{k,t,n,x_{tn}}$ (7) where $k \in \{1,\ldots,5\}$ in this experiment. Small probability values are omitted for brevity; however, a complete list in each row would of course sum to one. We can observe that some of the more common words share a number of topic-aspects which explain them with a certain probability.

In addition we show how documents can be augmented with terms suggested by the phantom. Table 3 lists the terms t for which $q_{k,t,n,x_{tn}}$ is the largest for the phantom aspect k in a document n. The results are given for ten randomly selected documents in the corpus. The terms are not present in the corresponding document; however, they fit nicely to the topical content of the original document, suggesting the possible use of this method for query expansion, as queries are typically short and incomplete.

Table 2 Analysis of three heterogeneous newsgroup documents

Words	Latent aspects and their posterior probabilities
system	*medical* 0.55, *cryptogr.* 0.44, *space* 0.01
medicin	*medical* 1.00
peopl	*religious* 0.75, *cryptogr.* 0.08, *medical* 0.13, *space* 0.04
public	*cryptogr.* 0.58, *religious* 0.42
system	*cryptogr.* 0.44, *medical* 0.19, *space* 0.37
agre	*religious* 0.95
faith	*religious* 1.00
accept	*religious* 0.88
christ	*religious* 1.00
teach	*religious* 0.97
clinic	*medical* 1.00
mission	*space* 1.00
religious	*religious* 1.00
jesu	*religious* 1.00
holi	*religious* 1.00
doctrin	*religious* 1.00
scriptur	*religious* 1.00
govern	*cryptogr.* 1.00
peopl	*cryptogr.* 0.66, *medical* 0.13, *space* 0.20
christ	*religious* 1.00
food	*medical* 1.00
rutger	*cryptogr.* 1.00
church	*religious* 1.00
atho	*religious* 1.00

The first column lists the words t which are present in the document n. In the second column, the most probable aspects k are given along with their posterior probabilities $q_{k,t,n,x_{tn}}$. Note the uncertainty in explaining some of the more common words

The above analyses cannot be computed for MB, LPCA, NMF or PLSA because no single component accounts for the missing terms.

3.6 Detecting and removing "false absences" and "false presences": an evaluation

In this section, we measure the performance of the AB model in detecting non-content-bearing causes. It should be stressed that there is no "clean" data available for training, instead the algorithm only sees the possibly corrupted data, without knowing about the existence of noise processes a priori.

Note that the use of factor models for noise separation and removal is not recent. PCA and ICA have both been used for this purpose quite extensively, in continuous-

Table 3 Expansion of ten randomly selected documents from the four newsgroups collection

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 escrow 0.83
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 0.96 comput 0.94
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
system medicin
effect 0.97 medic 0.96 peopl 0.96 doctor 0.92 patient 0.92 diseas 0.91 treatment 0.91 physician 0.89
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 med 0.24 diet 0.24
peopl effect doctor
medic 0.98 patient 0.87 diseas 0.85 treatment 0.84 medicin 0.84 physician 0.81 food 0.81
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

For each document, the first line contains the terms present in the document, followed by the top list of terms that the phantom-topic is responsible for, along with the posterior probability $q_{k,t,n,x_{in}}$ of the phantom

valued signal processing. Denoising of gene expression arrays [55] and denoising EEG signals [56] are two of the most known examples.

However, noise removal from discrete domains has not been attempted so far, up to the best of our knowledge, and the use of discrete factor models or aspect models to this task has not been studied. This is what we attempt in the remainder of this section, in the binary data setting.

Interestingly, in the natural binary data considered, we only encounter noise factors that create attribute absences by turning a 1 into a 0. Nevertheless, in order to show that our model is not restricted to detect this type of noise factor but also the symmetrical counterpart of it (when some of the zeros are randomly flipped to ones), we will create such situations artificially in some of the presented examples.

3.6.1 *Detection of missing or added remains from palaeontological data*

3.6.1.1 Filling in false absences We assume that a genus (an attribute) is absent in a site (an observation) either because the genus did not live in the area, or because it did but no remains were recorded. The former is a true absence and the latter a false absence. Reasons for the missingness were discussed in Sect. 3.1.1. One might quite safely assume that in case a genus is observed at several sites, the sites should be consecutive in their ages. That is, observing a genus at sites n and $n + l$, $l > 1$ implies that the genus should also have been observed at all intermediate sites $n + 1,\ldots,n + l - 1$, if the sites are sorted according to their ages. Not observing the genus t at an intermediate site n' means that the zero at $x_{n',t}$ is a false absence.

In the experiments that follow, the original data is fed to the AB model, without labels indicating the type of zeros. We would like to stress that the order of the observations is by no means utilised in the AB model or in the estimation procedure.

As the missingness is largely identified by one latent aspect as shown in Sect. 3.4, we can correct for the missingness by *post-processing* the data by removing the "phantom" aspect and reconstructing the data again. More precisely, first identify the phantom aspect by looking at the values of a_{tk} and finding the k for which $a_{tk} \approx 0\ \forall t$; denote this by k^*. Then remove the phantom aspect k^* by setting $s_{k^*n} = 0\ \forall n$ and normalise all s_{kn} such that $\sum_k s_{kn} = 1$ holds again for all n. Then compute the reconstruction of the data as $p_{tn} = \sum_k a_{tk}s_{kn}$ where s_{kn} was updated as described above. Finally, round the p_{tn} to 0 or 1.

For comparison, we also reconstruct the data by other methods: MB, LPCA, NMF and PLSA. In MB, LPCA and NMF, the reconstruction is computed similarly by rounding p_{tn} to 0 or 1, except that no component is removed, as the missingness in these methods is not separated by any one component but instead the components collaborate in explaining the data as it is. NMF and PLSA are not designed for binary data and are thus somewhat problematic to employ, due to the lack of suitable probabilistic interpretation. With NMF, values above 1 are possible as NMF does not treat the values as probabilities, so we

Table 4 Decrease in the number of missing values when the palaeontological data are reconstructed using the model parameters

AB post-proc.	AB	MB	LPCA	NMF	PLSA
745	47	54	155	75	−39

Generation of new missing values is possible, as indicated by the negative decrease of PLSA. "AB post-proc." refers to post-processing the data by removing the phantom

simply turn those to 1. The data model of PLSA is quite different too, as already discussed in Sect. 2.4. The parameters give $p(t|n)$, the probability of generating word t into any word position of document n having L_n words. In the palaeontological setting, "words" now correspond to genera, and "documents" correspond to sites. We resort to interpreting a 0–1 vector in the light of PLSA as follows. Let L_n be the unknown length of document n, and compute the probability of word t appearing at least once in the document—this corresponds to binary coding of the document. The probability is then

$$p(\text{'attribute } t \text{ appears at least once in observation } n') = 1 - (1 - p(t|n))^{L_n} \tag{32}$$

in which we assume the unknown document length L_n to be the number of ones in the observation.[6] The probability thus obtained is again rounded to 0 or 1.

Table 4 shows the decrease in the number of missing presences (false absences) when the data are reconstructed using AB, MB, LPCA, NMF and PLSA. The decrease is largest in AB when the data are post-processed by removing the phantom as described above; the result of plain AB without the post-processing is also given for comparison. It is well possible that new false absences are generated in the reconstruction process, if new 1s are inserted outside the original range of the observations of a genus. Indeed, such new missing values are generated especially at PLSA. The results shown are optimal among 50 random initialisations.

3.6.1.2 Detecting added noise A more challenging setting is obtained by artificially introducing an extra noise factor by randomly adding extra presences (1s) into the original data. In this case, not only the 0s have two underlying explanations (a "true absence" or a "false absence") but also the 1s may be true or false. We corrupt the data such that the proportion of extra[7] 1s in each observation (site) is distributed according to Uniform[0,0.4]; in the original data the percentage of 1s is 5.08% and in the corrupted data it is 12.5%—more than doubled. We then estimate $K = 5$ latent aspects in the corrupted data and obtain one "white phantom" having a negligible probability of generating any genus, and one "black phantom" having a large probability of generating any genus, and three real aspects.

The posterior probability $q_{k,t,n,x_{tn}}$ that the aspect k has generated the observation x_{tn} is computed as in Formula (7). The histograms of the posteriors $q_{k,t,n,x_{tn}}$ for true 1s, added 1s and 0s are seen in Fig. 8. They are computed as

$$p(k|\text{true ones}) \propto \sum_{t,n:x_{tn}=1 \text{ originally}} q_{k,t,n,x_{tn}} \tag{33}$$

$$p(k|\text{added ones}) \propto \sum_{t,n:x_{tn}=1 \text{ added}} q_{k,t,n,x_{tn}} \tag{34}$$

$$p(k|\text{zeros}) \propto \sum_{t,n:x_{tn}=0} q_{k,t,n,x_{tn}} \tag{35}$$

The quantities (33)–(35) are normalised such that each of them sums to 1 over k. We can see in Fig. 8a that the "white phantom" (the leftmost bar in all plots) has a very small or zero probability in true or added 1s and correspondingly a high probability at zeros. The "black phantom" (the third bar in all plots) has a large posterior probability in the added 1s and a very small probability at zeros. The real aspects behave in an opposite manner.

For comparison, Fig. 8b, c give the corresponding values for MB and PLSA. The number of components is chosen such that the total number of parameters is equal in all models considered—this gives $K = 19$ for MB and $K = 5$ for PLSA. At each model, the parameters used are from an in-sample log likelihood-optimal run over 10 repeated runs. No latent component differentiates between 0s and true and added 1s either in MB or in PLSA. Using $K = 5$ in MB would not result in a "white phantom" or a "black phantom" either. For LPCA and NMF, the quantities (33)–(35) cannot directly be computed, as the posterior of a component is not a well defined concept.

3.6.2 Detecting and correcting distortions in raster images

The data set of raster images of handwritten digits originally has no inherent pixel omissions or additions; therefore, it can be used for objective and controlled assessment. We create the two types of distortion studied in

[6] Another way would be to average over L_n, assuming that L_n ranges uniformly between the number of ones in the observation and some manually chosen upper limit; in Table 4 this would give inferior results for many choices of the upper limit.

[7] This is indeed not the proportion of 0s turned to 1s, but instead includes new 1s superimposed at existing 1s, which have no effect.

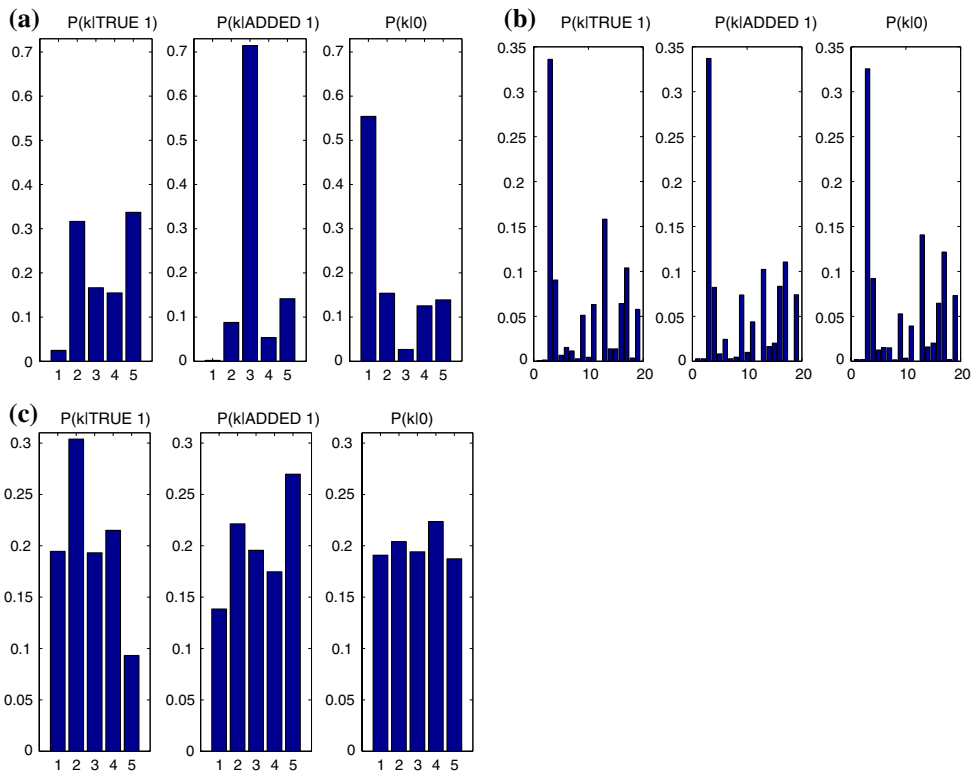

Fig. 8 Posterior probabilities of the latent aspects k in corrupted palaeontological data. **a** AB, **b** MB, **c** PLSA. At each case, the leftmost plot shows the probabilities at true 1s (33), the middle one at added 1s (34) and the rightmost at 0s (35). The number of components is chosen such that the total number of parameters is the same in all models. In AB, aspects $k = 1$ and 3 differentiate between the three cases

this section artificially and measure the ability of AB in detecting them.

First we add a corrosion cause into the data: we turn "off" a uniformly varying amount of pixels that were "on" in the original images. In the original data, any pixel that is "off" (0) is a "true absence" and can be explained by the content of the image. In the corrupted data, however, a 0 is either a true absence as before, or a false absence, explained by the corrosion.

3.6.2.1 Noise removal We then demonstrate the use of the AB model in noise removal. As the noise is identified by one latent aspect, we can correct for the noise by removing the noise aspect and reconstructing the data again. Similarly to what was described in Sect. 3.6.1, we identify k^* as the aspect corresponding to noise, by $a_{tk^*} \approx 0 \; \forall \; t$.[8] We then set $s_{k^* \, n} = 0 \; \forall \; n$ and normalise all s_{kn} by requiring $\sum_k s_{kn} = 1$. The reconstruction of the data is then computed by rounding $p_{tn} = \sum_k a_{tk} s_{kn}$ to 0 or 1.

Figure 9 shows the success in reconstructing corrupted digits where some pixels are turned to 0: the proportion of extra 0s is drawn from a Uniform[0,0.4] distribution. The noise removal rate is measured as $1 - (fp + fn)/2$ where fp is the rate of false positives, occurring if a true 0 is turned to 1, and fn is the rate of false negatives, occurring if a false 0 is not turned to 1. At MB, LPCA, NMF and PLSA, the reconstruction is computed as described in Sect. 3.6.1 related to Table 4.

In Fig. 9 we see that aspect Bernoulli is very successful in binary noise removal when the parameters are post-processed by removing the aspect corresponding to noise, as described above. Without such post-processing (not shown), AB behaves quite similarly to NMF. LPCA is comparable at very small K only, and PLSA is not very successful: in both methods, the rate of false negatives is quite large even though false positives are rare. The error bars give the standard error on both sides of the mean, over 5 disjoint subsets of the data.

3.6.2.2 Multiple causes of presences and absences Let us then see how the basis images combine to reconstruct instances of observed digit images. As an example we analyse the corrupted digit data where the proportion of extra 0s was drawn from a Uniform [0,0.4] distribution; the same data set was used to create Fig. 9. The number of latent aspects was chosen based on the AIC as $K = 14$. The top row of Fig. 10 shows the 14 bases (parameters a_{tk}) obtained for this data set; these are in-sample log likelihood optimal values over 30 random initialisations. In addition to bases that look like prototypical images as they contain high probabilities on corresponding pixels, we also have

[8] At large K, several aspects may correspond to noise, but for simplicity we only select the one having the smallest value of $\sum_t a_{tk}$.

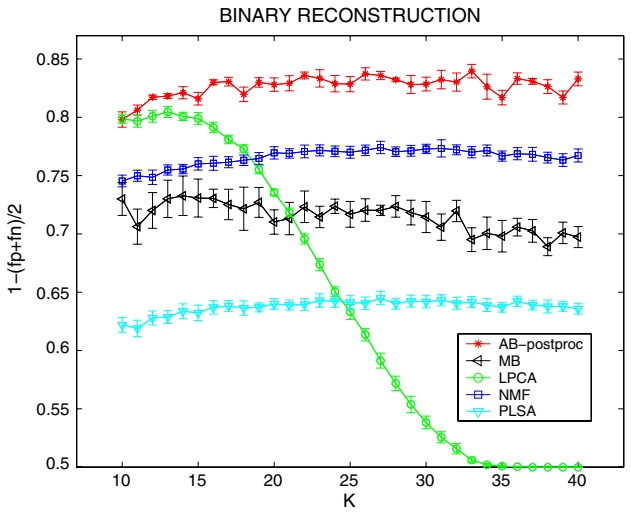

Fig. 9 Noise removal in artificially corrupted binary handwritten digit images. The parameters of AB are post-processed by removing the component explaining the noise, re-normalising the parameters, and reconstructing the data. In the other methods, the data are reconstructed from the original estimated parameters, as no single component explains the noise. *Horizontal axis* Number of components. *Vertical axis* $1 - (fp + fn)/2$ where $fp =$ false-positive rate, $fn =$ false-negative rate

one phantom basis for which a_{tk} is almost zero at all pixels t. To demonstrate the role of the phantom and the way the aspects may combine, we then analyse 6 observed images, shown in the leftmost column. For each image n and aspect k, the posterior probability q_{k,t,n,x_m} that the kth aspect explains the observed value (0 or 1) of all pixels $t = 1,\ldots,240$ is then given. On all these plots, the level of darkness of a pixel is proportional to the probability of it being "on".

The '5' depicted on the first data instance (second row of Fig. 10) is largely explained by the basis image which is a prototype of '5'. In addition, the basis '6' explains the pixels that are left unexplained by the basis '5'. A similar phenomenon is seen in the second and third data instances where a '6' and an '8' are analysed. The pixels that are "on" have multiple causes and so several bases contribute to explaining the observed data.

The fourth data instance is a '2' that has suffered corrosion. It is well explained by the basis '2', except for the pixels which are off due to the artificially created corrosion. These pixels are explained by the phantom with the highest probability. A similar case is seen in the fifth data instance where a corrupted '1' is analysed.

The last example does not directly resemble any one of the basis images, and it is explained by a combination of bases '7' and '6' and the phantom.

The bases given by MB, LPCA, PLSA and NMF are shown in Fig. 11. No single basis corresponds to the corruption, instead the bases resemble parts of digits. For the

Fig. 10 Results on artificially corrupted binary handwritten digit images where some pixels have been turned to white. The images on the *top line* depict the reshaped parameters a_{tk} as basis images. Some examples from this data set are shown in the first column, and their analysis as provided by the AB model in the next columns. For each datum instance n and each aspect k, the probability values q_{k,t,n,x_m} are shown for each pixel $t \in \{1,\ldots,240\}$. On all these plots, the level of darkness of a pixel is proportional to the probability of it being 'on'

ease of comparison, $K = 14$ bases are estimated; however, the results at different K are quite similar.

Similarly, using AB in the case of added 1s (not shown) we get one "black phantom" which has a high posterior probability of having created the non-content-bearing black pixels. The content-bearing pixels (both white and black) are explained by one or a few content-bearing latent aspects.

3.6.3 Detecting added words from Usenet text messages

In Sect. 3.5 we have seen that term occurrences in text messages naturally contain a factor of word omission and AB is able to detect that factor. However, for text, the goodness of this detection can only be assessed in a subjective manner. In order to conduct an objective evaluation we create an artificial setting, in the same way as with the palaeontological data. We randomly add 1s in the data such that the proportion of extra 1s in each document is distributed according to Uniform[0,0.4]; in the original data the percentage of 1s is 6.3% and in the added data it is 13.7%. We then estimate $K = 6$ latent aspects in the corrupted data, and obtain four aspects reflecting the four

Fig. 11 Basis images estimated by MB (*top row*), LPCA (*second row*), PLSA (*third row*) and NMF (*bottom row*) in artificially corrupted binary handwritten digits

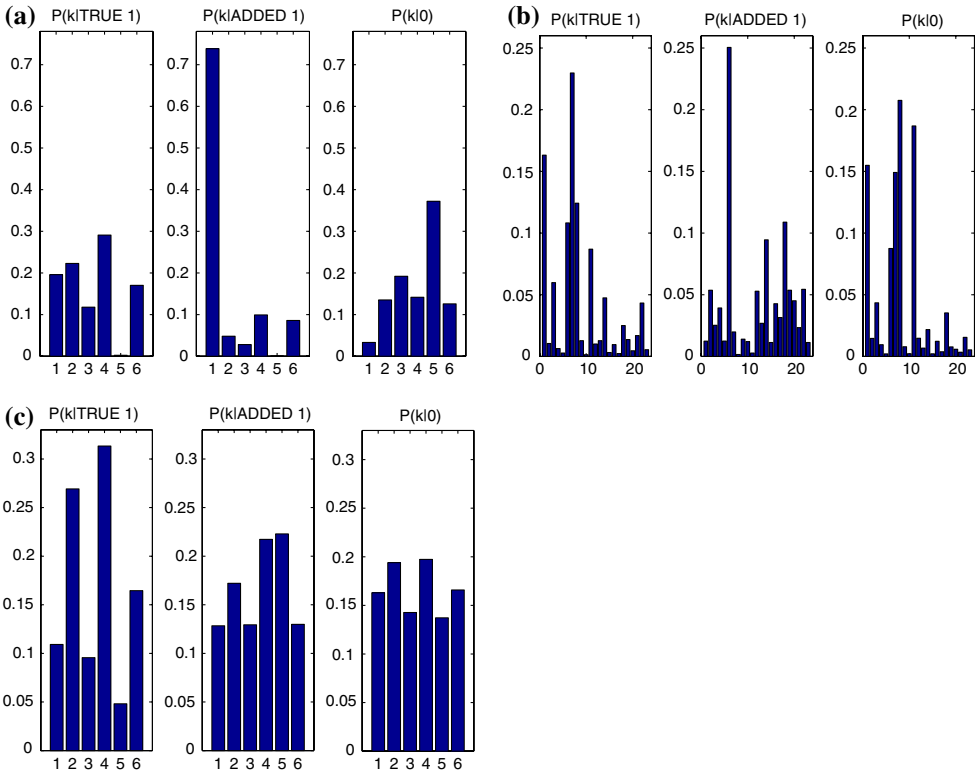

Fig. 12 Posterior probabilities of the latent aspects $k = 1,\ldots,6$ in corrupted newsgroup data. **a** AB, **b** MB, **c** PLSA. At each case, the leftmost plot shows the probabilities at true 1s (33), the middle one at added 1s (34) and the rightmost at 0s (35). In AB, aspects $k = 1$ and 5 differentiate between the three cases

newsgroups; in addition there is a "white phantom" having a negligible probability of generating any term, and a "black phantom" having a large probability of generating any term. The black phantom explains the artificially added terms which do not fit the topical contents of the documents.

We can also measure the degree to which the models are able to distinguish between different 1 s. Figure 12 shows the normalised histograms of the posterior probabilities of latent aspects k. For each k, we compute $p(k|\text{true } 1)$, $p(k|\text{added } 1)$ and $p(k|0)$ similarly as before by AB (a), MB (b) and PLSA (c). At MB and PLSA, the number of components is chosen such that the number of parameters in all methods are equal. The results shown are in-sample log likelihood optimal results over 10 random initialisations. In Fig. 12 (a) in the first histogram we see that the "white phantom" ($k = 5$) of AB explains none of the true 1s. Correspondingly, the "black phantom" ($k = 1$) explains the added 1s to a high degree, as seen in the second histogram. The third histogram shows that the white phantom ($k = 5$) explains most of the zeros whereas the black phantom ($k = 1$) only explains a small fraction of them. In text document data, the zeros might be "true absences" or "false absences" but we cannot manually distinguish between them, and so the numerical accuracies cannot be measured in this respect. In Fig. 12b, the sixth Bernoulli mixture component explains the added 1s to a high degree, but it also explains the true 1s and 0s to a large degree. In Fig. 12c, none of the PLSA components deviates.

4 Conclusions

This paper presented a probabilistic multiple cause model for 0–1 data. The AB model analyses the causes behind not only the presences (1) but also the absences (0) of attributes, and produces interpretable explanations to these, which is in contrast to all existing models for 0–1 data. A distinctive feature of the aspect Bernoulli model is its ability to separate binary noise factors (both omissions and additions) in the data by automatically creating specific "phantom" latent aspects. A "white phantom" gives a negligible probability of appearance to any attribute and thus it is used to explain omissions in the data; in contrast, a "black phantom" generates occurrences of all attributes with probability close to 1 and as such it explains additions in the data. The phantoms are not hard-coded into the model but arise automatically.

We have also demonstrated how the AB model outperforms related models in the task of noise removal from binary data. In addition we studied and contrasted AB to related Bernoulli models in several settings in terms of scaling, out-of-sample likelihood and parameter interpretability: AB scales equally to the mixtures of Bernoulli

model and outperforms that in terms of out-of-sample likelihood; AB scales favourably compared with logistic PCA while their out-of-sample likelihoods tend to be similar; finally, AB gives interpretable parameters whereas logistic PCA does not.

In addition to the variety of related factorisation models discussed, let us briefly mention a few models that hard-wire the presence of a common (noise-)component. The mixture of Gaussians model of Law et al. [57] has one content-bearing latent cause for each observation; then for each attribute, the value of the attribute is either generated from a distribution specific to the latent cause chosen, or from a common cause. The models of Hofmann [58], Barnard et al. [59, 60] and Blei et al. [61] present hierarchical architectures where the latent components are arranged into a tree; the root node is a common component that may participate in the generation of all observations. Recently a somewhat similar tree-construction has been considered explicitly for finding uninformative features by Wang and Kabán [62].

An intermediate model between Logistic PCA and aspect Bernoulli could also be constructed for completeness. The likelihood of such a model reads $p(\boldsymbol{x}|\boldsymbol{s},\boldsymbol{a}) = \prod_n \prod_t g\left(\sum_k a_{tk}s_{kn}\right)^{x_{tn}}\left(1 - g\left(\sum_k a_{tk}s_{kn}\right)\right)^{1-x_{tn}}$ where the parameters a_{tk} and s_{kn} are not restricted to probabilities. In our studies (not shown), using $g(u) = (\exp(u) - 1)/(\exp(u) + 1)$, the results of such a model have indeed consistently been between those of LPCA and AB in all respects. However, the data representation is similar to NMF, and the noise is not separated out into any specific components.

In this paper, we have shown how the AB model can successfully analyse both noisy and noiseless 0–1 data in a variety of application areas, of which the palaeontological setting is perhaps the most demanding. From a palaeontologist's point of view, the possibility to distinguish between true and false absences has great appeal, as there are several systematic and random sources of bias in the data collection process. In addition to studies involving palaeobiodiversity and turnover, the method has potential applicability in palaeoecology, including the generation of "proxy" data for palaeoenvironment reconstruction, for palaeocommunity reconstruction, and for the study of evolutionary dynamics at the community and metacommunity levels. A very practical use of the method is to characterise and summarise the taxonomic deficiencies of the palaeontological data: for example, a group of genera (attributes) having a lot of false absences can be concluded as too noisy to be included in further studies.

Acknowledgments The authors wish to thank Professor Heikki Mannila for insightful discussions regarding the model and the palaeontological data, and for suggesting Fig. 7. The authors would also like to thank the anonymous reviewers for their useful comments and suggestions.

Appendix A

The following holds for any distributions $q_n(\cdot)$:

$$\sum_n \log p(\boldsymbol{x}_n|\boldsymbol{s}_n,\boldsymbol{a}) = \sum_n \sum_{z_n} q_n(z_n) \log p(\boldsymbol{x}_n|\boldsymbol{s}_n,\boldsymbol{a}) \tag{36}$$

$$= \sum_n \sum_{z_n} q_n(z_n) \log \frac{p(\boldsymbol{x}_n, z_n|\boldsymbol{s}_n,\boldsymbol{a})}{p(z_n|\boldsymbol{x}_n,\boldsymbol{s}_n,\boldsymbol{a})} \frac{q_n(z_n)}{q_n(z_n)} \tag{37}$$

$$= \underbrace{\sum_n \left[\sum_{z_n} q_n(z_n) \log p(\boldsymbol{x}_n, z_n|\boldsymbol{s}_n,\boldsymbol{a}) - \sum_{z_n} q_n(z_n) \log q_n(z_n)\right]}_{F_{s_1,\ldots,s_N,\boldsymbol{a}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\, q_1(\cdot),\ldots,q_N(\cdot))} + \sum_n \underbrace{\sum_{z_n} q_n(z_n) \log \frac{q_n(z_n)}{p_n(z_n|\boldsymbol{x}_n,\boldsymbol{s}_n,\boldsymbol{a})}}_{KL(q_n(\cdot)||p(\cdot|\boldsymbol{x}_n,\boldsymbol{s}_n,\boldsymbol{a}))} \tag{38}$$

We can recognise that the first term of F in (38) is the sum of conditional expectations of the joint likelihood of the data and latent variables $\boldsymbol{z}_n$, conditioned on $\boldsymbol{s}_n$. The second term is the sum of entropies of q_n. Finally, the last term is the sum of Kullback–Leibler distances between the (so far arbitrary) distributions q_n over $\boldsymbol{z}_n$ and the true conditional posterior distributions of $\boldsymbol{z}_n$, conditioned on $\boldsymbol{s}_n$.

Since the KL divergence is always positive [63] (which can easily be shown by using Jensen's inequality), F is always a lower bound to the log likelihood $\log p(\boldsymbol{x}_n|\boldsymbol{s}_n,\boldsymbol{a})$, irrespective of the distributions $q_n(\cdot)$. When $q_n(\cdot) \equiv p(\cdot|\boldsymbol{x}_n,\boldsymbol{s}_n,\boldsymbol{a})$, $\forall n = 1,\ldots,N$, then the KL divergence becomes zero and, therefore, the lower bound approaches the log likelihood exactly.

The iterative EM procedure is then: In the E-step, having some fixed estimates of $\boldsymbol{s}_n$, $n = 1,\ldots,N$ and $\boldsymbol{a}$, we maximise $F_{s_1,\ldots,s_N,\boldsymbol{a}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\, q_1(\cdot),\ldots,q_N(\cdot))$ with respect to all $q_n(\cdot)$. This is achieved by setting these to the true conditional posteriors $p(\cdot|\boldsymbol{x}_n,\boldsymbol{s}_n,\boldsymbol{a})$, for all $n = 1,\ldots,N$, which—from Bayes' rule—are the following:

$$p(z_n|\boldsymbol{x}_n,\boldsymbol{s}_n,\boldsymbol{a}) = \frac{p(\boldsymbol{x}_n|z_n,\boldsymbol{a})p(z_n|\boldsymbol{s}_n)}{\sum_{z_n} p(\boldsymbol{x}_n|z_n,\boldsymbol{a})p(z_n|\boldsymbol{s}_n)} \tag{39}$$

$$= \frac{\prod_t \prod_k [a_{tk}^{x_{tn}}(1-a_{tk})^{1-x_{tn}}]^{z_{tnk}} \prod_t \prod_k [s_{kn}]^{z_{tnk}}}{\prod_t \sum_k s_{kn} a_{tk}^{x_{tn}}(1-a_{tk})^{1-x_{tn}}} \tag{40}$$

$$= \prod_t \frac{\prod_k [s_{kn} a_{tk}^{x_{tn}}(1-a_{tk})^{1-x_{tn}}]^{z_{tnk}}}{\sum_k s_{kn} a_{tk}^{x_{tn}}(1-a_{tk})^{1-x_{tn}}} \tag{41}$$

$$= \prod_t p(z_{tn}|x_{tn},\boldsymbol{s}_n,\boldsymbol{a}_t) \tag{42}$$

In (39), we used that $p(\boldsymbol{x}_n|\boldsymbol{z}_n,\boldsymbol{a}) = p(\boldsymbol{x}_n|\boldsymbol{s}_n,\boldsymbol{z}_n,\boldsymbol{a})$, which follows from the dependency structure of AB, namely that $\boldsymbol{x}_n$

depends on $\boldsymbol{s}_n$ only through $\boldsymbol{z}_n$, therefore, knowing $\boldsymbol{z}_n$ makes $\boldsymbol{x}_n$ independent of $\boldsymbol{s}_n$.

It may be interesting to note that the above conditional posterior distribution $p(\boldsymbol{z}_n|\boldsymbol{x}_n,\boldsymbol{s}_n,\boldsymbol{a})$ factorises naturally, without having imposed a factor form. Of course, as we know, this posterior is conditioned on the value of $\boldsymbol{s}_n$, so even though it is an exact conditional posterior quantity, there is no tractable exact posterior over the joint distribution of all hidden variables of the model, which would be $p(\boldsymbol{s}_n,\boldsymbol{z}_n|\boldsymbol{x}_n,\boldsymbol{a}) = p(\boldsymbol{s}_n|\boldsymbol{x}_n, \boldsymbol{a})p(\boldsymbol{z}_n|\boldsymbol{x}_n,\boldsymbol{s}_n, \boldsymbol{a})$. The latter term is what we just computed, while the former term is intractable as discussed earlier in the main text, and so a point estimate for $\boldsymbol{s}_n$ will be obtained as part of the M-step.

In the M-step, we keep the posterior distributions $q_n(\cdot)$ fixed at the values computed in the previous E-step, and compute the most probable value of $\boldsymbol{s}_n$ for each $\boldsymbol{x}_n$ as well as the parameters $\boldsymbol{a}$. This is achieved by maximising $F_{s_1,\ldots,s_N,\boldsymbol{a}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\ q_1(\cdot),\ldots,q_N(\cdot))$ with respect to all $\boldsymbol{s}_n$ and $\boldsymbol{a}$.

To ensure the constraint $\sum_k s_{kn} = 1$ is met, we add a Lagrangian term. Denoting

$$q_{k,t,n,x_{tn}} \equiv p(z_{tn} = k|x_{tn},\boldsymbol{s}_n,\boldsymbol{a}_t) = \frac{s_{kn}a_{tk}^{x_{tn}}(1-a_{tk})^{1-x_{tn}}}{\sum_\ell s_{\ell n}a_{t\ell}^{x_{tn}}(1-a_{t\ell})^{1-x_{tn}}} \tag{43}$$

where the last equality follows from (41), and replacing the result obtained from the previous E-step into F, the expression to maximise, up to a constant term, is

$$Q = \sum_n \left[\sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n)\log p(\boldsymbol{x}_n,\boldsymbol{z}_n|\boldsymbol{s}_n,\boldsymbol{a}) - \lambda_n\left(\sum_k s_{kn} - 1\right)\right] \tag{44}$$

$$= \sum_n \left[\sum_t\sum_k \left[\sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) z_{tnk}\right] \times \log[s_{kn}a_{tk}^{x_{tn}}(1-a_{tk})^{1-x_{tn}}] - \lambda_n\left(\sum_k s_{kn} - 1\right)\right] \tag{45}$$

$$= \sum_n \left[\sum_t\sum_k q_{k,t,n,x_{tn}} \log[s_{kn}a_{tk}^{x_{tn}}(1-a_{tk})^{1-x_{tn}}] - \lambda_n\left(\sum_k s_{kn} - 1\right)\right] \tag{46}$$

where λ_n are Lagrange multipliers, and the second term of F (the entropy term) was omitted for being a constant w.r.t. the variables of interest. Eq. (45) was obtained by expanding $p(\boldsymbol{x}_n,\boldsymbol{z}_n|\boldsymbol{s}_n,\boldsymbol{a}) = \prod_t\prod_k[s_{kn}a_{tk}^{x_{tn}}(1-a_{tk})^{1-x_{tn}}]^{z_{tnk}}$ and grouping together the terms with z_{tnk}. To obtain (46), we used the result (42) and the notation (43), so that $\sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n)z_{tnk} = \sum_{\boldsymbol{z}_n}\prod_t p(z_{tn}|x_{tn},\boldsymbol{s}_n,\boldsymbol{a}_t)z_{tnk} = p(z_{tn} = k|x_{tn}, \boldsymbol{s}_n,\boldsymbol{a}_t) = q_{k,t,n,x_{tn}}$.

The terms that depend on elements of $\boldsymbol{s}_n$ are

$$Q_{\boldsymbol{s}_n} = \sum_k\sum_t q_{k,t,n,x_{tn}}\log s_{kn} - \lambda_n\left(\sum_k s_{kn} - 1\right) + \text{const.} \tag{47}$$

Now, we solve the system of stationary equations w.r.t. s_{kn}, which are the following.

$$\frac{\partial Q_{\boldsymbol{s}_n}}{\partial s_{kn}} = \sum_t q_{k,t,n,x_{tn}}/s_{kn} = \lambda_n \tag{48}$$

Multiplying both sides by s_{kn}, we obtain

$$\sum_t q_{k,t,n,x_{tn}} = \lambda_n s_{kn} \tag{49}$$

from which we have that

$$s_{kn} = \sum_t q_{k,t,n,x_{tn}}/\lambda_n \tag{50}$$

The value of λ_n is obtained by summing both sides, and using $\sum_k s_{kn} = 1$. This gives $\lambda_n = \sum_k\sum_t q_{k,t,n,x_{tn}} = T$, since by its definition, $\sum_k q_{k,t,n,x_{tn}} = 1$.

To complete the M-step, we now maximise Q w.r.t. $\boldsymbol{a}$. The terms that depend on elements of $\boldsymbol{a}$ up to constants, are the following.

$$Q_{\boldsymbol{a}} = \sum_n\sum_k\sum_t q_{k,t,n,x_{tn}}[x_{tn}\log a_{tk} + (1-x_{tn}) \times \log(1-a_{tk})] \tag{51}$$

The stationary equations are then the following.

$$\frac{\partial Q_{\boldsymbol{a}}}{\partial a_{tk}} = \sum_n q_{k,t,n,x_{tn}}\left(\frac{x_{tn}}{a_{tk}} - \frac{1-x_{tn}}{1-a_{tk}}\right) = \sum_n q_{k,t,n,x_{tn}}\frac{x_{tn}-a_{tk}}{a_{tk}(1-a_{tk})} = 0 \tag{52}$$

The denominator is the variance of the Bernoulli and always non-negative, it can be simplified and by isolating a_{tk} we have the solution:

$$a_{tk} = \frac{\sum_n x_{tn}q_{k,t,n,x_{tn}}}{\sum_n q_{k,t,n,x_{tn}}} \tag{53}$$

Note that the constraint $a_{tk} \in [0,1]$ needed not be explicitly imposed in this model setting, as it is automatically satisfied for binary data.[9]

[9] This also fallows from the first moment identity for exponential family of distributions, of which the Bernoulli distribution is a member [50].

Appendix B

The derivation of the fixed point equations (13)–(15) as an alternating optimisation of $\sum_n p(\boldsymbol{x}_n|\, \boldsymbol{s}_n,\boldsymbol{a})$ (or equivalently $\sum_n p(\boldsymbol{x}_n, \boldsymbol{s}_n|\boldsymbol{a})$) is as follows.

Denote $\overline{a_{tk}} = 1 - a_{tk}$. The log likelihood (11) is maximised, subject to the constraints $\sum_k s_{kn} = 1$ and $a_{tk} + \overline{a_{tk}} = 1$. The corresponding Lagrangian is thus the following:

$$\mathcal{L} = \sum_n \sum_t \left[x_{tn} \log \sum_k a_{tk} s_{kn} + (1 - x_{tn}) \log \sum_k \overline{a_{tk}} s_{kn} - c_{tk}(a_{tk} + \overline{a_{tk}} - 1) - \lambda_n \left(\sum_k s_{kn} - 1 \right) \right] \tag{54}$$

where c_{tk} and λ_n are Lagrangian multipliers, and we have rewritten $(1 - \sum_k a_{tk} s_{kn})$ as $\sum_k \overline{a_{tk}} s_{kn}$. The stationary equations of $\mathcal{L}$ with respect to both a_{tk} and $\overline{a_{tk}}$ are

$$\frac{\partial \mathcal{L}}{\partial a_{tk}} = \sum_n \frac{x_{tn}}{\sum_\ell a_{t\ell} s_{\ell n}} s_{kn} - c_{tk} = 0 \tag{55}$$

$$\frac{\partial \mathcal{L}}{\partial \overline{a_{tk}}} = \sum_n \frac{1 - x_{tn}}{\sum_\ell \overline{a_{t\ell}} s_{\ell n}} s_{kn} - c_{tk} = 0 \tag{56}$$

Multiplying the first of the above equations by a_{tk} and the second by $\overline{a_{tk}}$ we obtain

$$a_{tk} \sum_n \frac{x_{tn}}{\sum_\ell a_{t\ell} s_{\ell n}} s_{kn} - c_{tk} a_{tk} = 0 \tag{57}$$

$$\overline{a_{tk}} \sum_n \frac{1 - x_{tn}}{\sum_\ell \overline{a_{t\ell}} s_{\ell n}} s_{kn} - c_{tk} \overline{a_{tk}} = 0 \tag{58}$$

Summing both sides and using $a_{tk} + \overline{a_{tk}} = 1$ provides the Lagrangian multiplier c_{tk}:

$$c_{tk} = a_{tk} \sum_n \frac{x_{tn}}{\sum_\ell a_{t\ell} s_{\ell n}} s_{kn} + \overline{a_{tk}} \sum_n \frac{1 - x_{tn}}{\sum_\ell \overline{a_{t\ell}} s_{\ell n}} s_{kn} \tag{59}$$

as in (15). From (57) we have the solution for a_{tk} in the form of a fixed point equation:

$$a_{tk} = a_{tk} \sum_n \frac{x_{tn}}{\sum_\ell a_{t\ell} s_{\ell n}} s_{kn} / c_{tk} \tag{60}$$

as in (14). Solving for s_{kn} proceeds similarly: the stationary equation is

$$\frac{\partial \mathcal{L}}{\partial s_{kn}} = \sum_t \left(\frac{x_{tn}}{\sum_\ell a_{t\ell} s_{\ell n}} a_{tk} + \frac{1 - x_{tn}}{\sum_\ell \overline{a_{t\ell}} s_{\ell n}} \overline{a_{tk}} \right) - \lambda_n = 0 \tag{61}$$

Multiplying both sides by s_{kn} we obtain

$$\sum_t \left(\frac{x_{tn}}{\sum_\ell a_{t\ell} s_{\ell n}} a_{tk} s_{kn} + \frac{1 - x_{tn}}{\sum_\ell \overline{a_{t\ell}} s_{\ell n}} \overline{a_{tk}} s_{kn} \right) = \lambda_n s_{kn} \tag{62}$$

Summing over k and using $\sum_k s_{kn} = 1$ we have the Lagrange multiplier λ_n:

$$\lambda_n = \sum_t \frac{x_{tn}}{\sum_\ell a_{t\ell} s_{\ell n}} \sum_k a_{tk} s_{kn} + \sum_t \frac{1 - x_{tn}}{\sum_\ell \overline{a_{t\ell}} s_{\ell n}} \sum_k \overline{a_{tk}} s_{kn} \tag{63}$$

$$= \sum_t x_{tn} + \sum_t (1 - x_{tn}) = T \tag{64}$$

Having computed λ_n, from (62) we obtain the fixed point equation for s_{kn}, identical to (13):

$$s_{kn} = s_{kn} \left\{ \sum_t \frac{x_{tn}}{\sum_\ell a_{t\ell} s_{\ell n}} a_{tk} + \frac{1 - x_{tn}}{\sum_\ell \overline{a_{t\ell}} s_{\ell n}} \overline{a_{tk}} \right\} / T \tag{65}$$

As discussed in the text, this derivation is simpler and yields the same multiplicative updates (13)–(15), obtained also via rewriting the EM algorithm (7)–(9). However, the fact that we need not iterate each multiplicative fixed point update to convergence separately before alternating these inner loops, but we can actually just alternate them while still obtainining a convergent algorithm, is less apparent from the derivation given in this section. Instead, this is a consequence of the EM interpretation presented in Appendix A. Indeed, recall that every single multiplicative update is a combination of a full E-step and an M-step update for one (group of) variables, hence it is guaranteed not to decrease the likelihood.

References

1. Harman HH (1967) Modern factor analysis, 2nd edn. University of Chicago Press, Chicago
2. Saund E (1995) A multiple cause mixture model for unsupervised learning. Neural Comput 7:51–71
3. Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 42:177–196
4. Hyvärinen A, Karhunen J, Oja E (2001) Independent component analysis. Wiley Interscience, New York
5. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
6. Kabán A, Bingham E, Hirsimäki T (2004) Learning to read between the lines: the aspect Bernoulli model. In: Proceedings of the 4th SIAM international conference on data mining, pp 462–466
7. Buntine W (2002) Variational extensions to EM and multinomial PCA. In: Machine learning: ECML 2002. Lecture notes in artificial intelligence (LNAI), vol 2430. Springer, New York, pp 23–34
8. McCallum A, Nigam K (1998) A comparison of event models for naive Bayes text classification. In: Sahami M (ed) Learning for text categorization. Papers from the AAAI Workshop, Technical report WS-98-05, AAAI, pp 41–48
9. Fortelius M, Werdelin L, Andrews P, Bernor RL, Gentry A, Humphrey L, Mittmann W, Viranta S (1996) Provinciality, diversity, turnover and paleoecology in land mammal faunas of the later Miocene of western Eurasia. In: Bernor R, Fahlbusch V, Mittmann W (eds) The evolution of Western Eurasian Neogene Mammal Faunas. Columbia University Press, Columbia, pp 414–448
10. Srebro N (2004) Learning with matrix factorizations. Ph.D. thesis, Massachusetts Institute of Technology

11. Hofmann T, Puzicha J (1998) Unsupervised learning from dyadic data. Technical Report TR-98-042, International Computer Science Insitute, Berkeley
12. Hofmann T (2004) Latent semantic models for collaborative filtering. ACM Trans Inf Syst 22:89–115
13. Marlin B, Zemel R (2004) The multiple multiplicative factor model for collaborative filtering. ICML-2004. In: Proceedings of the 21st international conference on machine learning, pp 576–583
14. Buntine W, Jakulin A (2006) Discrete components analysis. In: Saunders C, Grobelnik M, Gunn S, Shawe-Taylor J (eds) Subspace, latent structure and feature selection techniques. Springer, Berlin
15. Marlin B (2004) Modeling user rating profiles for collaborative filtering. In: Thrun S, Saul L, Schölkopf B (eds) Advances in neural information processing systems, vol 16. MIT, Cambridge
16. Kabán A, Bingham E (2006) ICA-based binary feature construction. In: Rosca JP, Erdogmus D, Príncipe JC, Haykin S (eds) Independent component analysis and blind signal separation. 6th international conference, ICA 2006, Proceedings. Lecture notes in computer science, vol 3889. Springer, Berlin, pp 140–148
17. Kabán A, Bingham E. Factorisation and denoising of 0–1 data: a variational approach. Neurocomputing, special issue on advances in blind signal processing (in press)
18. Everitt BS, Hand DJ (1981) Finite mixture distributions. Chapman & Hall, London
19. Gyllenberg M, Koski T, Reilink E, Verlaan M (1994) Non-uniqueness in probabilistic numerical identification of bacteria. J Appl Probab 31:542–548
20. Meilă M (1999) An accelerated Chow and Liu algorithm: fitting tree distributions to high-dimensional sparse data. In: ICML '99: Proceedings of the sixteenth international conference on machine learning, Morgan Kaufmann, San Francisco, pp 249–257
21. Schein A, Saul L, Ungar L (2003) A generalized linear model for principal component analysis of binary data. In: Bishop CM, Frey BJ (eds) Proceedings of the 9th international workshop on artificial intelligence and statistics
22. Tipping ME (1999) Probabilistic visualisation of high-dimensional binary data. In: Kearns MS, Solla SA, Cohn DA (eds) Advances in neural information processing systems, vol 11, pp 592–598
23. Collins M, Dasgupta S, Schapire RE (2001) A generalization of principal component analysis to the exponential family. Advances in neural information processing systems, vol 14, pp 617–624
24. Hofmann T, Puzicha J (1999) Latent class models for collaborative filtering. In: Dean T (ed) Proceedings of the 16th international joint conference on artificial intelligence, IJCAI 99, Morgan Kaufmann, San Francisco, pp 688–693
25. Dayan P, Zemel RS (1995) Competition and multiple cause models. Neural Comput 7:565–579
26. Seppänen JK, Bingham E, Mannila H (2003) A simple algorithm for topic identification in 0–1 data. In: Lavrač N, Gamberger D, Todorovski L, Blockeel H (eds) Knowledge discovery in databases: PKDD 2003. Lecture notes in artificial intelligence, vol 2832, Springer, New York, pp 423–434
27. Jaakkola TS (1997) Variational methods for inference and estimation in graphical models. Ph.D. thesis, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA
28. Barlow H, Kaushal T, Mitchison G (1989) Finding minimum entropy codes. Neural Comput 1:412–423
29. Földiák P (1990) Forming sparse representations by local anti-Hebbian learning. Biol Cybern 64:165–170
30. Schmidhuber J (1992) Learning factorial codes by predictability minimization. Neural Comput 4:863–879
31. Zemel RS (1993) A minimum description length framework for unsupervised learning. Ph.D. thesis, University of Toronto
32. Kemp C, Griffiths TL, Tenenbaum JB (2004) Discovering latent classes in relational data. Technical Report AI Memo 2004-019, Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory
33. Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo AI (1996) Fast discovery of association rules. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds) Advances in knowledge discovery and data mining, Chap. 12. AAAI, New York, pp 307–328
34. Dhillon IS (2001) Co-clustering documents and words using bipartite spectral graph partitioning. Technical Report TR 2001-05, Department of Computer Sciences, University of Texas, Austin
35. Smolensky P (1986) Information processing in dynamical systems: foundations of harmony theory. In: Rumelhart DE, McClelland JL (eds) Parallel distributed processing: explorations in the microstructure of cognition, Foundations, vol 1. MIT, Cambridge, pp 194–281
36. Jolliffe IT (1986) Principal component analysis. Springer, Berlin
37. Paatero P, Tapper U (1994) Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5:111–126
38. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791
39. Lee DD, Seung HS (2000) Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems, vol 13, 556–562
40. Tipping ME, Bishop CM (1999) Probabilistic principal component analysis. J R Stat Soc B Stat Methodol 61:611–622
41. Dahyot R, Charbonnier P, Heitz F (2004) A bayesian approach to object detection using probabilistic appearance-based models. Pattern Anal Appl 7:317–332
42. Srebro N, Jaakkola T (2003) Weighted low-rank approximations. In: Fawcett T, Mishra N (eds) Machine learning. Proceedings of the twentieth international conference (ICML 2003). AAAI, New York, pp 720–727
43. Gordon GJ (2003) Generalized^2 linear^2 models. In: Becker S, Thrun S, Obermayer K (eds) Advances in neural information processing systems, vol 15. MIT, Cambridge, pp 577–584
44. Haft M, Hofman R, Tresp V (2003) Generative binary codes. Pattern Anal Appl 6:269–284
45. Barry JC, Morgan M, Flynn L, Pilbeam D, Behrensmeyer A, Raza S, Khan I, Badgley C, Hicks J, Kelley J (2002) Faunal and environmental change in the late Miocene Siwaliks of Northern Pakistan. Palaeobiology (Supplement). 28:1–71
46. Puolamäki K, Fortelius M, Mannila H (2006) Seriation in paleontological data using Markov chain Monte Carlo methods. PLoS Comput Biol 2:e6
47. Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge
48. Smyth P (2000) Model selection for probabilistic clustering using cross-validated likelihood. Stat Comput 10:63–72
49. Heckerman D, Chickering DM (1996) A comparison of scientific and engineering criteria for Bayesian model selection. Technical Report MSR-TR-96-12, Microsoft Research
50. Bernardo JM, Smith AFM (1994) Bayesian theory. Wiley, New York
51. Akaike H (1973) Information theory and an extension of the maximum likelihod principle. In: Petrox B, Csaki F (eds) Second international symposium on information theory, pp 267–281
52. Kabán A (2007) Predictive modelling of heterogeneous sequence collections by topographic ordering of histories. Mach Learn 68:63–95

53. Peterson C, Söderberg B (1989) A new method for mapping optimization problems onto neural networks. Int J Neural Syst 1:3–22
54. Wu J-M, Chiu S-J (2001) Independent component analysis using Potts models. IEEE Trans Neural Netw 12:202–211
55. Jung T-P, Makeig S, Humphries C, Lee T-W, McKeown MJ, Iragui V, Sejnowski TJ (2000) Removing electroencephalographic artifacts by blind source separation. Psychophysiology 37:163–178
56. Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. PNAS 97:10101–10106
57. Law MHC, Figueiredo MAT, Jain AK (2004) Simultaneous feature selection and clustering using mixture models. IEEE Trans Pattern Anal Mach Intell 26:1154–1166
58. Hofmann T (1999) The cluster-abstraction model: unsupervised learning of topic hierarchies from text data. In: Dean T (ed) Proceedings of the 16th international joint conference on artificial intelligence, IJCAI 99. Morgan Kaufmann, San Francisco, pp 682–687
59. Barnard K, Duygulu P, Forsyth D (2001) Clustering art. IEEE Conf Comput Vis Pattern Recognit 2(II):434–441
60. Barnard K, Duygulu P, Forsyth D, de Freitas N, Blei DM, Jordan MI (2003) Matching words and pictures. J Mach Learn Res 3:1107–1135
61. Blei DM, Griffiths TL, Jordan MI, Tenenbaum JB (2004) Hierarchical topic models and the nested Chinese restaurant process. In: Thrun S, Saul L, Schölkopf B (eds) Advances in neural information processing systems, vol 16. MIT, Cambridge
62. Wang X, Kabán A (2005) Finding uninformative features in binary data. In: Gallagher M, Hogan JM, Maire F (eds) Intelligent data engineering and automated learning—IDEAL 2005. Lecture notes in computer science, vol 3578. Springer, New York, pp 40–47
63. Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York

Author Biographies

Ella Bingham received her M.Sc. degree in Engineering Physics and Mathematics at Helsinki University of Technology in 1998, and her Dr.Sc. degree in Computer Science at Helsinki University of Technology in 2003. She is currently at Helsinki Institute for Information Technology, located at the University of Helsinki. Her research interests include statistical data analysis and machine learning.

Ata Kabán is a lecturer in the School of Computer Science of the University of Birmingham, since 2003. She holds a B.Sc. degree in computer science (1999) from the University "Babes-Bolya" of Cluj-Napoca, Romania, and a Ph.D. in computer science (2001) from the University of Paisley, UK. Her current research interests concern statistical machine learning and data mining. Prior to her career in computer science, she obtained a B.A. degree in musical composition (1994) and the M.A. (1995) and Ph.D. (1999) degrees in musicology from the Music Academy "Gh. Dima" of Cluj-Napoca, Romania.

Mikael Fortelius is a palaeontologist with special interest in plant-eating mammals of the Cenozoic, especially ungulates and their relationship with habitat and climate change (the Ungulate Condition). Mikael is Professor of Evolutionary Palaeontology in the Department of Geology and Group Leader in the Institute of Biotechnology (BI), University of Helsinki. Since 1992, he has been engaged in developing a database of Neogene Old World Mammals (http://www.helsinki.fi/science/now/). The NOW database is maintained at the Finnish Museum of Natural History and developed in collaboration with an extensive Advisory Board; data access and downloading are entirely public.