The Aspect Bernoulli Model: Multiple Causes of Presences and Absences

Ella Bingham[*] ELLA@IKI.FI
HIIT Basic Research Unit, Department of Computer Science, University of Helsinki, Finland
and
Neural Networks Research Centre, Helsinki University of Technology, Finland

Ata Kabán A.KABAN@CS.BHAM.AC.UK
School of Computer Science, University of Birmingham, UK

Mikael Fortelius MIKAEL.FORTELIUS@HELSINKI.FI
Division of Palaeontology, University of Helsinki, Finland

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. The observations consist of remains of mammal genera found at different sites of excavation. In this data, two types of zeros arise naturally: a zero, indicating that remains of a genus were not observed at a particular site, arises either because the genus did not live in the site, or because it did but no remains were found. The former is a true absence and the latter a false absence. 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. Specifically, by being a factor model, the AB model is also a powerful probability estimator for high dimensional data analysis problems and outperforms related models in terms of generalisation ability in the case of the data sets considered.

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 bio-informatics, 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 or hidden causes and their combinations. Revealing these causes gives new insight into the underlying

[*] Part of the work performed while visiting the School of Computer Science, University of Birmingham, UK.

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 (Harman, 1967; Saund, 1995; Hofmann, 2001; Hyvärinen et al., 2001; Blei et al., 2003, 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. Some multiple cause models also allow for multi-way clustering of data, meaning that both observations and attributes can be clustered according to which latent causes they are due to.

We present a probabilistic, multiple-cause latent variable model for binary data. The Aspect Bernoulli (AB) model, previously presented in a short preliminary version (Kabán et al., 2004), can formally be seen as a Bernoulli analogue of the multinomial decomposition model known under the names of aspect model, PLSA (Hofmann, 2001), and their generative versions such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Multinomial Principal Component Analysis (MPCA) (Buntine, 2002). 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)} (as in McCallum & Nigam, 1998). A distinctive feature of the AB model is that noise in this event space is explicitly modelled by specialised components, and this may further be straightforwardly exploited for noise removal from binary data.

Multiple-cause models for binary data have been devised before in the literature. Most notably, Saund's model (Saund, 1995) asserts an interaction model for the 1s in the data, which takes the form of a noisy-OR. However, the 0-s are suppressed this way, and observing 0-s remains a default uninteresting event. By contrast, in a 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 are left 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).

In addition to the mentioned potential uses, in this paper, we demonstrate the abilities of the Aspect Bernoulli model in an actual application, in the context of palaeontological data (Fortelius et al., 1996) 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? This is the sort of question we try to answer under the AB generative modelling assumption. Indeed, 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; we will discuss this in more detail in Section 3.1.1. 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 Aspect Bernoulli 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 (2004). It can also be seen as a special case of the URP model of Marlin and Zemel (2004), if the observations were to be restricted to 0/1. Further, closely related models were discussed by Hofmann et al. (1998, 2004) and others. A more complete review of related models will be given in Section 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 factorisation model for the specific purpose of reasoning about 0-1 data by detecting and removing "noise" factors. There is no readily available algorithm for this task. Secondly, it is also of interest here to study how such a specific instantiation of factorisation models compares to other models 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 Section 2. Experimental results are shown in Section 3, demonstrating and Section 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 Aspect Bernoulli (AB) model, and derive the algorithm for estimating the model parameters. We then discuss some particular features of the model, such as its scaling and its ability to detect omissions and additions. We end by contrasting the AB model to various other multiple cause models.

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 denote the observations, attributes and latent aspects, respectively. Let $\mathbf{x}_n$ denote a T-dimensional multivariate binary observation and x_{tn} the value of its t-th attribute. The elements x_{tn} may be generated by different latent aspects k with probabilities specific to the observation and aspect in question. The n-th observation vector $\mathbf{x}_n$ 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 $\mathbf{x}_n$, the following two steps are taken:
  - Pick a latent aspect k with probability $P(k|n)$
  - Let the latent aspect k generate a 1 (presence) or a 0 (absence) of the t-th attribute. The Bernoulli probability of generating 1, $P(1|k,t)$, only depends on k and t and is not specific to 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 t-th 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 t-th attribute of the n-th 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 $z_{tnk} = \delta(z_{tn} - k)$. The conditional complete data likelihood now reads as

$$p(\mathbf{x}_n, \mathbf{z}_n|\mathbf{s}_n) = \prod_{t=1}^{T}\prod_{k=1}^{K}[s_{kn}a_{tk}^{x_{tn}}(1-a_{tk})^{1-x_{tn}}]^{z_{tnk}} \tag{1}$$

where $\mathbf{s}_n = (s_{1n}, \ldots, s_{Kn})$, are the probabilities of selecting one of the K aspects and $\mathbf{z}_n = (z_{1n1}, \ldots, z_{Tn1}, \ldots, z_{TnK})$. The model assumes that the elements x_{tn} of $\mathbf{x}_n$ are conditionally independent given the latent variable $\mathbf{z}_n$. 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.

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

The expected complete data log likelihood is

$$E\{\mathcal{L}^{\rm c}\} = E\{\log p(\mathbf{x}_1,\ldots,\mathbf{x}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N|\mathbf{s}_1,\ldots,\mathbf{s}_N)\} = \sum_{n=1}^{N} E\{\log p(\mathbf{x}_n,\mathbf{z}_n|\mathbf{s}_n)\}$$

$$= \sum_{n,t,k} q_{k,t,n,x_{tn}} \log[s_{kn} a_{tk}^{x_{tn}} (1-a_{tk})^{1-x_{tn}}] \tag{2}$$

$$= \sum_{n,t,k} q_{k,n,t,x_{tn}} [\log s_{kn} + x_{tn} \log a_{tk} + (1-x_{tn}) \log(1-a_{tk})] \tag{3}$$

where $q_{k,n,t,x_{tn}} = E_{P(z_{tn}|t,n,x_{tn})}[z_{tnk}]$ is evaluated as the posterior probability that the aspect k has generated the observation (either the 0 or the 1) at the t-th attribute of the n-th instance.

The maximisation of the expected complete data log likelihood (3) — by taking derivatives, equating to zero and also taking into account the constraint $\sum_k s_{kn} = 1$ with the aid of Lagrange multipliers — then leads to the iterative EM algorithm below.

$$q_{k,t,n,x_{tn}} = \frac{s_{kn} a_{tk}^{x_{tn}} (1-a_{tk})^{1-x_{tn}}}{\sum_\ell s_{\ell n} a_{t\ell}^{x_{tn}} (1-a_{t\ell})^{1-x_{tn}}} \tag{4}$$

$$s_{kn} = \sum_t q_{k,t,n,x_{tn}}/T \tag{5}$$

$$a_{tk} = \frac{\sum_n x_{tn} q_{k,t,n,x_{tn}}}{\sum_n q_{k,t,n,x_{tn}}} \tag{6}$$

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 t-th element of any observation vector $\mathbf{x}_n$. A well-known extension of this is the single-cause mixtures of Bernoulli (MB) model (Everitt & Hand, 1981; Gyllenberg et al., 1994) $\sum_k \pi_k \prod_t a_{tk}^{x_{tn}} (1-a_{tk})^{1-x_{tn}}$ where $a_{tk} = P(1|t,k)$ and π_k is the prior probability of the k'th 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 Aspect Bernoulli model, and we see this by summing out $\mathbf{z}$ from (1):

$$p(\mathbf{x}_n|\mathbf{s}_n) = \sum_{\mathbf{z}} p(\mathbf{x}_n,\mathbf{z}_n|\mathbf{s}_n) = \sum_{\mathbf{z}} \prod_{t=1}^{T} \prod_{k=1}^{K} [s_{kn} a_{tk}^{x_{tn}} (1-a_{tk})^{1-x_{tn}}]^{z_{tnk}}$$

$$= \prod_{t=1}^{T} \sum_{k=1}^{K} s_{kn} a_{tk}^{x_{tn}} (1-a_{tk})^{1-x_{tn}} \tag{7}$$

$$= \prod_{t=1}^{T} (\sum_{k=1}^{K} a_{tk} s_{kn})^{x_{tn}} (1 - \sum_k a_{tk} s_{kn})^{1-x_{tn}}. \tag{8}$$

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 the z_{tnk} equals 1 at each pair t and n.

Finally, to see the equivalence between (7) and (8), notice that when $x_{tn} = 1$, then according to both (7) and (8) we have that $p(x_{tn}|\mathbf{s}_n) = \sum_k a_{tk}s_{kn}$; and for the case when $x_{tn} = 0$ we have $p(x_{tn}|\mathbf{s}_n) = \sum_k (1 - a_{tk})s_{kn}$ from both (7) and (8). 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 (8) indeed resembles the well-known Bernoulli likelihood if we denote by $p_{tn} := p(x_{tn} = 1|\mathbf{s}_n) = \sum_k a_{tk}s_{kn}$ 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 binary data, as will be seen in the experimental section.

We can also rewrite the (4-6) in order to gain savings in the memory requirements and obtain a convenient matrix-form notation. The derivation of this iterative algorithm is given in the Appendix.

$$s_{kn} = s_{kn}\{\sum_t \frac{x_{tn}}{\sum_\ell a_{t\ell}s_{\ell n}} a_{tk} + \frac{1 - x_{tn}}{1 - \sum_\ell a_{t\ell}s_{\ell n}}(1 - a_{tk})\}/T \tag{9}$$

$$a_{tk} = a_{tk}\sum_n \frac{x_{tn}}{\sum_\ell a_{t\ell}s_{\ell n}} s_{kn}/c_{tk} \tag{10}$$

where the denominator evaluates as

$$c_{tk} = a_{tk}\sum_n \frac{x_{tn}}{\sum_\ell a_{t\ell}s_{\ell n}} s_{kn} + (1 - a_{tk})\sum_n \frac{(1 - x_{tn})}{1 - \sum_\ell a_{t\ell}s_{\ell n}} s_{kn}. \tag{11}$$

Finally, the summary of this fixed point algorithm in matrix-form notation is listed below.

- Initialise $\mathbf{A}$ and $\mathbf{S}$ within the appropriate domains.
- Iterate until convergence

$$\bar{\mathbf{A}} = 1 - \mathbf{A} \tag{12}$$

$$\mathbf{S} = \mathbf{S} \otimes \{\mathbf{A}^{\mathbf{T}}[\mathbf{X} \oslash \mathbf{AS}] + \bar{\mathbf{A}}^{\mathbf{T}}\left[(\mathbf{1} - \mathbf{X}) \oslash \bar{\mathbf{A}}\mathbf{S}\right]\} \tag{13}$$

$$\mathbf{S} = \mathbf{S} \oslash \mathbf{Z} \tag{14}$$

$$\mathbf{A} = \mathbf{A} \otimes \{[\mathbf{X} \oslash \mathbf{AS}]\,\mathbf{S}^{\mathbf{T}}\} \tag{15}$$

$$\bar{\mathbf{A}} = \bar{\mathbf{A}} \otimes \{\left[(\mathbf{1} - \mathbf{X}) \oslash \bar{\mathbf{A}}\mathbf{S}\right]\mathbf{S}^T\} \tag{16}$$

$$\mathbf{A} = \mathbf{A} \oslash \left(\mathbf{A} + \bar{\mathbf{A}}\right) \tag{17}$$

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

Note that only those elements of the product $\mathbf{AS}$ need to be computed for which we have a nonzero observation element in the corresponding position. Likewise, only those elements of the product $\bar{\mathbf{A}}\mathbf{S} = \mathbf{1} - \mathbf{AS}$ need to be computed for which the corresponding element in the data matrix is zero. Sparsity of any of the quotient matrices can be then taken advantage of in the matrix multiplications in the algorithm.

2.2 Scaling

The scaling per iteration of the ML estimation of an Aspect Bernoulli is $\mathcal{O}(NTK)$. This is less convenient as the $\mathcal{O}(\#(nonzero)\,K)$ 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 as our Bernoulli model operates on the event space of attribute values, proposing to explain both the presences and the absences of each attribute and each datum instance. By contrast, multinomial models operate on the event space of attributes, and are only concerned with explaining the attribute occurrences. Bernoulli component models, with very few exceptions (Meilă, 1999), typically do not scale better than AB: The scaling per iteration of the Bernoulli mixtures is the same $\mathcal{O}(NTK)$. Logistic PCA (Schein et al., 2003), 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.

2.3 Omission/addition detection by "phantom" latent aspects

Let us now discuss the ability of the Aspect Bernoulli 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 automatically accomplished by additional 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 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. It can be shown that the Aspect Bernoulli algorithm minimises the weighted sum of entropies of all its parameters as a by-product of the likelihood maximisation. This implies that if there is a process that generates non-content-bearing zeros or ones in the data, it will be detected and modelled by the phantom aspects. The entropy argument is seen in the following: Rewrite the expected complete data likelihood (3) by the help of formulae (5-6) and obtain the following.

$$
\begin{aligned}
E\{\mathcal{L}^{\mathrm{c}}\} &= \sum_{n,k} \log s_{kn} \sum_{t} q_{k,n,t,x_{tn}} + \sum_{k,t} \log a_{tk} \sum_{n} x_{tn} q_{k,n,t,x_{tn}} + \\
&\quad + \sum_{k,t} \log(1-a_{tk}) \sum_{n} (1-x_{tn}) q_{k,n,t,x_{tn}} \qquad (18) \\
&= T \sum_{n,k} s_{kn} \log s_{kn} + \sum_{k,t} a_{tk} \log a_{tk} \sum_{n} q_{k,n,t,x_{tn}} + \\
&\quad + \sum_{k,t} (1-a_{tk}) \log(1-a_{tk}) \sum_{n} q_{k,n,t,x_{tn}} \qquad (19) \\
&= -T \sum_{n} H(\mathbf{s}_n) - \sum_{k,t} H([a_{tk}, 1-a_{tk}]) \sum_{n} q_{k,n,t,x_{tn}} \qquad (20)
\end{aligned}
$$

Here $\mathbf{s}_n$ refers to the distribution $(s_{1n}, \ldots, s_{Kn})$. Now each parameter a_{tk} and $\mathbf{s}_n$ is a function of $q_{k,t,n,x_{tn}}$. Consequently each E-step will necessarily increase (20). Thus, the iterative EM maximisation of the model likelihood could also be interpreted as an iterative minimisation of a weighted sum of entropies. This implies two representational tendencies of the model:

- a tendency towards a sparse distribution of $\mathbf{s}_n$ (only a few latent aspects are active at a time), due to the first sum of terms in (20)
- a tendency towards extreme binary values in a_{tk}, due to the second sum of terms in (20)

Specifically, in the extreme case when the data supports 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 (20), so in such a case 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 $\mathbf{s}_n$ when more than one s_{kn} become active for a given n.

Interestingly, quite a similar analysis 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.

In the experiments that follow we will demonstrate detailed experimental analyses of the behaviour of “white” and “black” phantoms in the Aspect Bernoulli model.

2.4 Relation to other models

Starting from the factorisation in (8), 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 (Everitt & Hand, 1981; Gyllenberg et al., 1994), already mentioned in Section 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 (Schein et al., 2003) and the models of Tipping (1999) and Collins et al. (2001) 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 $p_{tn} = 1/(1 + e^{-\theta_{tn}})$. 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, and also possibly over-fitting. In contrast, the parameters of the linear decomposition in the Aspect Bernoulli 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, Hofmann, 2001), LDA (Latent Dirichlet allocation, Blei et al., 2003) and MPCA (Multinomial PCA, Buntine, 2002) models for multinomial data have been quite popular over the last few years. These factorise the multinomial parameter vector into a convex combination of latent causes. The PLSA is "aspect multinomial", in which the statistical events are the attributes, and the decomposition reads $p(t|n) = \sum_k a_{tk} s_{kn}$. In contrast, in the Aspect Bernoulli model the statistical events are the outcomes of the attributes (presences and absences) and these are the events that we seek to reason about.

As already mentioned, AB could formally be seen as a special case of the URP model (Marlin, 2004). The URP model was designed for collaborative filtering, and it assumes that the ratings or attribute outcomes are natural numbers in a given range. Each attribute is modelled as a multinomial variable over the possible outcomes. Thus, with ratings restricted to 0/1, URP reduces to AB — however, such a model has not been previously analysed.

Another collaborative filtering method (Polčicová & Tiňo, 2004), one that could be termed "aspect binomial", assumes that there is an inherent order scale of the attribute outcomes. No such ordering is imposed either in the aspect multinomial (PLSA) or Aspect Bernoulli (AB) models — the attributes of PLSA are not ordered, and similarly the outcomes 0 and 1 of AB could be replaced with any binary values. Collaborative filtering models for binary data have also been studied by Hofmann and Puzicha (1999) and Hofmann (2004); the model presented in the latter can be used for arbitrary response scales.

Saund's model (1995) 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 (Saund, 1995) and a mean-field approximate solution has been provided later (Dayan & Zemel, 1995). A somewhat similar model is the topic model presented by Seppänen et al. (2003); 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 (Seppänen et al., 2003). A discrete logical OR function is also discussed by Jaakkola (1997) who gives upper and lower bounds for the likelihood.

Early approaches to multiple cause models have been presented by Barlow et al. (1989), Földiák (1990), Schmidhuber (1992) and Zemel (1993); in these models the data is not necessarily assumed binary valued. Later, Dayan and Zemel (1995) 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 (2004) 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 (2004).

Non-probabilistic methods for the analysis of binary data include the method of frequent sets (Agrawal et al., 1996) 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 (Dhillon, 2001), 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 binary data is the famous Boltzmann machine (Smolensky, 1986).

Apart from binary data, well known methods for factoring continuous data include principal component analysis (PCA, Jolliffe, 1986), independent component analysis (ICA, Hyvärinen et al., 2001) and nonnegative matrix factorisation (NMF, Lee & Seung, 1999, 2000). 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.

Srebro and Jaakkola (2003, 2004) and Gordon (2003) discuss the general class of matrix factorisations and give an overall view to the problem.

As already stated, our inferential purposes are different from those of the mentioned works, since none of these formally related models have been used for inferring and removing non-content-bearing causes from the data. However, we placed our approach in the more general context of matrix factorisation and multiple cause modelling literature, and we now turn to experimentally demonstrate the use of AB on real world data sets, contrasting it to some of the related models reviewed here.

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 model parameters. Additional analyses are given in the end of this Section where the model's ability to detect omissions and additions is demonstrated.

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 collaboration between mammal palaeontologists[2]. The NOW data derive from the published literature as well as from unpublished compilations by contributors.

The dataset 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 degrees eastern longitude. Our dataset 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 per cent of the entries are 1. The data matrix is seen in Figure 1.

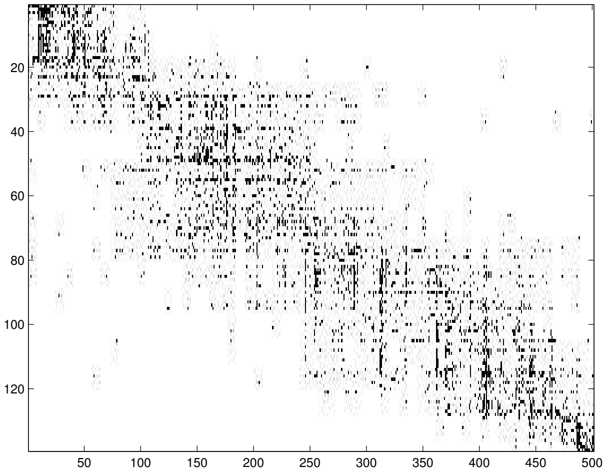


Figure 1: Palaeontological data. Rows correspond to genera and columns to sites of excavation.

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. (1996) or the NOW web site. The age estimates in our dataset 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

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

be recorded, and the number of rarer genera present continues to increase with sampling for most represented sample sizes (Fortelius et al., 1996). 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 are 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 of these phenomena incur absences of attributes in the data.

The NOW data used here are quite typical of palaeontological datasets; if anything, most datasets 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., Barry et al., 2002; Puolamäki et al., 2005). Our AB analysis may provide new insights into this issue, as will be demonstrated in the sequel.

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 digital 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 20Newsgroup corpus[4]: short Usenet messages from 4 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 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

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

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

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

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

The issue of model selection is now addressed, that is, how many components is the optimal choice. Indeed, the optimal model order may depend on the application (Ripley, 1996) and this is often overlooked in the machine learning literature; on the other hand, Smyth (2000) emphasises the use of cross validation. 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 (1996) in the Bayesian model selection framework. We will consider both cases 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 2 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 this data. The procedure we are using is known as "empirical Bayes test likelihood":

$$-\log \int_{\mathbf{s}_{test}} p(\mathbf{x}_{test}|\mathbf{s}_{test})p(\mathbf{s}_{test})d\mathbf{s}_{test} \tag{21}$$

where $p(\mathbf{s}_{test}) = 1/N_{train}\sum_n \delta(\mathbf{s}_{test} - \mathbf{s}_n)$ meaning that the vector $\mathbf{s}_{test}$ is sampled from the distribution of the estimates of $\mathbf{s}$ as obtained from the training data (Bernardo & Smith, 1994; Blei et al., 2003). For AB, the empirical Bayes test likelihood is computed as the following:

$$-\log \frac{1}{N}\sum_n \prod_t \left(\sum_k a_{tk}s_{kn}\right)^{x_{t,test}} \left(1 - \sum_k a_{tk}s_{kn}\right)^{1-x_{t,test}} \tag{22}$$

and for MB as

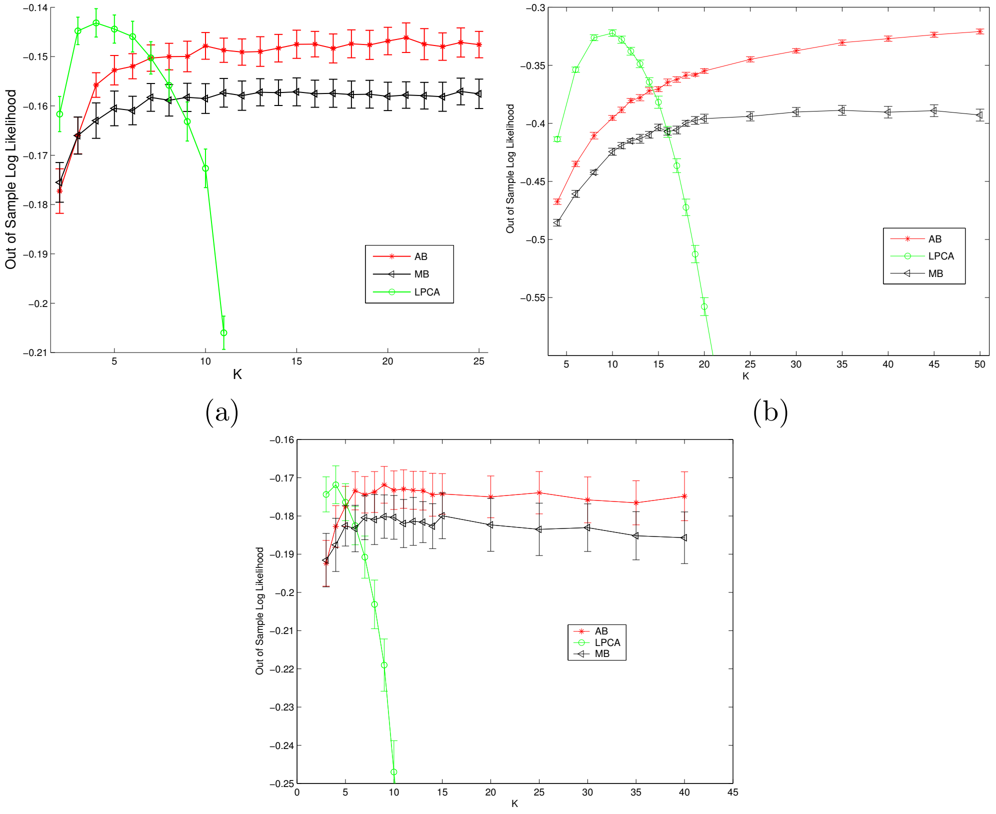
$$-\log \frac{1}{N}\sum_n \sum_k \pi_k \prod_t a_{tk}^{x_{t,test}}(1-a_{tk})^{1-x_{t,test}} \tag{23}$$

and for LPCA respectively

$$-\log \frac{1}{N}\sum_n \prod_t \left(\frac{1}{1+\exp(-\sum_k a_{tk}s_{kn})}\right)^{x_{t,test}} \left(1 - \frac{1}{1+\exp(-\sum_k a_{tk}s_{kn})}\right)^{1-x_{t,test}} . \tag{24}$$

Here n ranges over the training points; specifically, the s_{kn} are obtained for the training point $\mathbf{x}_n$. Instead, $x_{t,test}$ is the t-th dimension of a new, previously unseen test point.

The out of sample likelihoods of PLSA and NMF cannot be directly compared to AB in this setting, as they are not functions of strictly the same data: the zero entries do not contribute to the log likelihood as $\log(\sum_k a_{tk}s_{kn})^x = 0$ when $x = 0$. (The similarity of NMF and PLSA has been discussed by Buntine (2002).)



(c)

Figure 2: Out of sample log likelihood for LPCA, AB and MB, measuring the predictive capabilities of the models on the (a) palaeontological data, (b) handwritten digits data and (c) newsgroup document data. Horizontal axis: model order (number of estimated components K). The error bars show one standard error on both sides of the mean of the folds in 10-fold cross-validation.

From the results of Figure 2 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 as the data set is quite small). Also, the peak values of LPCA and AB are not significantly different: non-parametric Wilcoxon Rank-Sum tests on the best results (LPCA at 4 and AB at 20 components in the palaeontological data; LPCA at 10 and AB at 50 components in the digits data; and LPCA at 4 and AB at 9 components in newsgroup documents) show that the differences between these are not statistically significant at the 5% level: the P value is 0.52 in the palaeontological data, 0.62 in the digits data and 0.97 in the newsgroup data. Thus we can conclude that AB and LPCA are comparable in terms of prediction performance on these data.

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.) LPCA however over-fits badly after 4 components in the palaeontological and newsgroup data and after 10 components in the digit data.

AB may thus be a preferable choice for modelling and analysis because of AB's additional intuitive data explanatory capabilities, which will be demonstrated in the following Section. Additionally, the comparative computational scaling of these two models advantages AB over LPCA further, as discussed in Section 2.2.

3.2.2 AIC-based model selection for data explanation

In terms of data explanation, the AB model is of interest, as by construction it captures the generative process that reflects our intuition about the data. Contrarily to prediction tasks, we now seek to obtain a parsimonious data explanatory model. Following the arguments given by Ripley (1996), a procedure designed to achieve this goal is the Akaike Information Criterion (AIC) (Akaike, 1973). This has a very simple form as follows:

$$AIC(K) = -2\mathcal{L}(\mathcal{K}) + 2P(K) \tag{25}$$

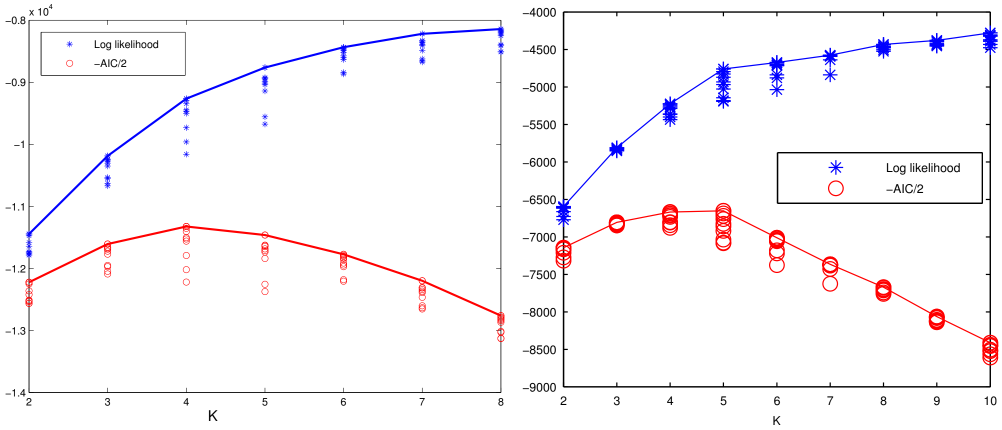
where $\mathcal{L}$ is the trained log likelihood of the model (must not be normalised i.e. it must not be divided by the size of the data), P is the number of free parameters that need to be estimated and the factor of 2 has historical reasons only. For the case of the Aspect Bernoulli model,

$$P(K) = TK + (K-1)N \tag{26}$$

where K is the assumed number of components. (The parameters s_{kn} sum to 1 over k, reducing the number of free parameters by one.) The optimal model order is then found by minimising (25) under K:

$$K_{opt} = \operatorname*{argmin}_{K} AIC(K) \tag{27}$$

The log likelihood and AIC-penalised log likelihood values in palaeontological data can be comparatively seen in Figure 3 (a) as obtained in 15 randomly (uniformly) initialised independent runs for each choice of K ranging from 2 to 8. The maximum log likelihood values have been selected for each value of K and these have then been used to create the model selection curve shown in the figure. (Log likelihoods smaller than the maximum are local optima.) Naturally, the log likelihood continues to increase with increasing model



(a) (b)

Figure 3: Log likelihoods ($*$) and AIC-penalised log likelihoods ($\circ$) of the AB model in repeated experiments on the (a) palaeontological data, (b) newsgroup data. The values are not normalised by the size of the data. Horizontal axis: model order (number of latent components K).

complexity — however, the AIC-penalised log likelihood peaks at $K = 4$ and thus this model order can be chosen.

Similarly, the AIC-optimal model order for the newsgroup data is $K = 5$ components, as seen in Figure 3 (b). In the handwritten digits data, the AIC suggests $K = 15$; the natural variation (i.e. several styles of 1-s, 4-s etc.) in handwriting styles requires more than 10 components to be used.

3.3 Parameter interpretability

Although our data sets are very different in their nature, in this Section we will demonstrate that the modelling assumptions of AB give rise to quite intuitive and interpretable representations on all three data sets analysed.

3.3.1 Clustering tendency and prototypical representations

As detailed in Section 2.3, in case the data contain natural clusters, the AB model's tendency to produce minimum entropy mixing proportions makes it suitable to finding these clusters. In this case, the components become cluster prototypes (cluster centres). To demonstrate this, we consider the set of images of handwritten numerals, which clearly has a clustered structure. The latent causes $\mathbf{a}_k = (a_{1k}, \ldots, a_{Tk})$ estimated from this data are shown in Figure 4 where the model order $K = 15$ has been chosen by the Akaike Information Criterion (25). White encodes zero and darker means higher probability. The causes indeed display prototypical images of the numerals. Quite a similar result is obtained by the Mixtures of Bernoulli model.

The cluster-based prototypical representation tendency will also be easily observed on text document representations. Moreover, text documents and the palaeontological data provide a more interesting setting, as they both naturally contain omitted presences of various attributes, as will be shown in the sequel.

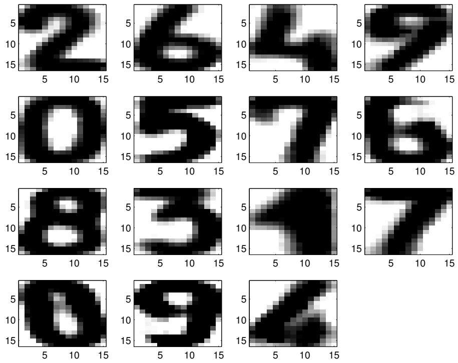


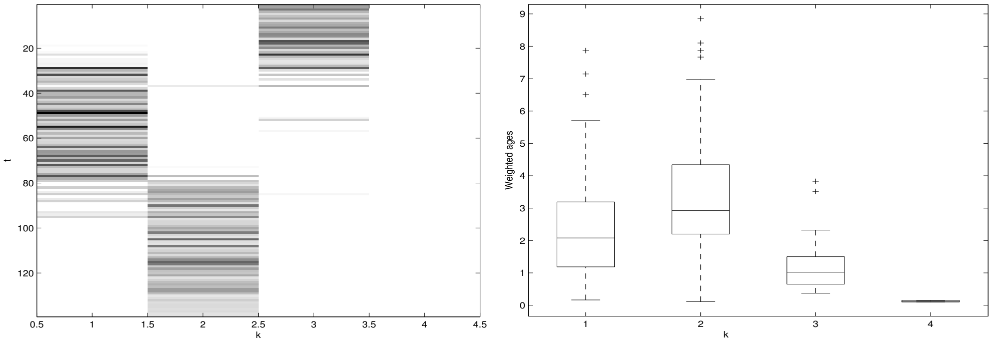
Figure 4: Latent causes in handwritten digits presented as prototype images given by the parameters a_{tk} in the Aspect Bernoulli model. $K = 15$ latent causes were estimated.

3.3.2 Representation of palaeontological data

Based on 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 Aspect Bernoulli model. Figure 5 (a) 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". As already mentioned, a white phantom explicitly generates zeros in the data, in contrast to other latent aspects that generate both zeros and ones.

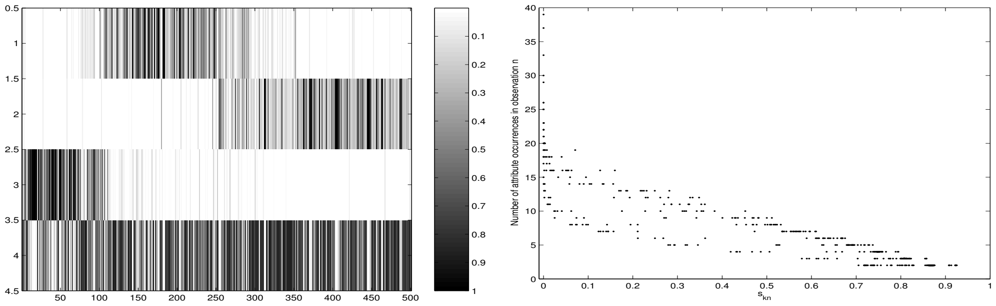
Let us visualise the grouping of genera by drawing a box plot of the ages of genera captured by different latent aspects. Figure 5 (b) shows for each latent aspect k the distribution of the ages of genera t weighted by the probabilities a_{tk}. We can see that different latent aspects indeed concentrate 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 Figure 6 (a). 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 of observations; by its presence, it generates absences of attributes, as seen in Figure 5 (a).



(a) (b)

Figure 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) $t = 1, \ldots, 139$. (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) (b)

Figure 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.

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 Figure 6 (b). At each observation, the probabilities of different aspects k need to sum to 1, and it is quite natural that observations having few 1s have small probabilities for aspects generating 1s. The Aspect Bernoulli model thus explains the 0s as well as the 1s, in contrast to most models which only explain the 1s and treat all zeros as a default value. In the view of the AB model, a "false absence" is not "part of the background" but a true entity which occurs due to a hidden cause and, as such, is modelled by the "phantom" aspect.

The parameters a_{tk} and s_{kn} given by the LPCA, MB, PLSA and NMF models (not shown) do not demonstrate a 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 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.

3.3.3 Text document representation

We now turn to the newsgroup document data and demonstrate the latent aspects found by the Aspect Bernoulli 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 (25). 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 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 k-th 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 notice 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 noticed 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

religious	*phantom*	*cryptographic*	*medical*	*space-related*
45.1	152.9	42.9	48.2	59.0
god 1.00	agre 1.3e-03	kei 1.00	effect 0.84	space 0.76
christian 1.00	sternlight 1.0e-11	encrypt 1.00	peopl 0.72	nasa 0.59
peopl 0.95	bless 3.2e-12	**system** 1.00	medic 0.66	orbit 0.49
rutger 0.81	truth 2.5e-15	govern 0.90	doctor 0.52	man 0.37
word 0.63	peopl 2.4e-15	public 0.89	patient 0.47	cost 0.35
church 0.63	comput 2.8e-16	clipper 0.84	diseas 0.42	**system** 0.34
bibl 0.61	system 8.6e-19	chip 0.83	treatment 0.40	pat 0.33
faith 0.60	man 1.1e-19	secur 0.82	medicin 0.40	launch 0.32
christ 0.59	nsa 1.0e-21	peopl 0.70	food 0.35	mission 0.30
jesu 0.56	shuttl 4.1e-22	comput 0.65	med 0.33	flight 0.28

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). 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 $\sum_n s_{kn}$ which reflects the overall probability of aspect k.

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.

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. Indeed, 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, and it is seen that the length and the topical content of the document effect the value of s_{kn}.

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 $P(k|n, t, x_{tn}) = q_{k,t,n,s_{tn}}$ (4) where $k \in \{1, ..., 5\}$ 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 $P(k|t, n, x_{tn})$ 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 a promising method of query expansion, as queries are typically short and incomplete.

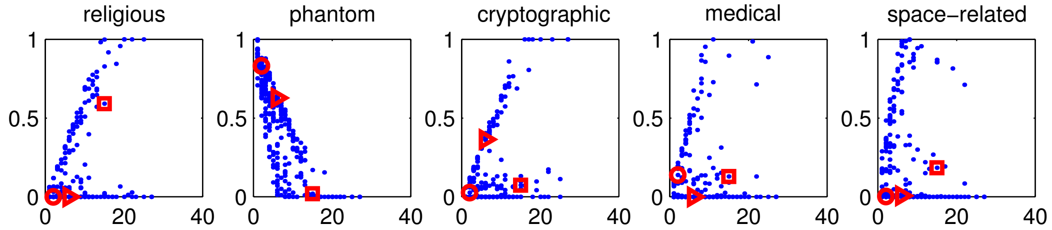


Figure 7: The probability s_{kn} of the individual aspects (topics) plotted against the number of different words in document n (horizontal axis). The probability of the phantom-topic (aspect no. 2) is negatively correlated with the richness of the documents, whereas the real-topics are positively correlated with the richness. Three documents are highlighted: $\circ$: 'system' 'medicin' — a very short document; $\Box$: '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 $\triangleright$: 'govern' 'secur' 'access' 'scheme' 'system' 'devic' — a medium size document focused on a single topic.

The above analyses can not be computed for the other models (MB, LPCA, NMF or PLSA) because no single component accounts for the missing terms.

3.4 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 apriori.

Interestingly, in the natural binary data considered, we only encounter noise factors that create attribute absences by turning a 1 into a 0. However, in order to show that our model is not restricted to detecting 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.

3.4.1 Detection of missing or added remains from palaeontological data

Filling in missing values. 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 Section 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

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

Table 2: Analysis of three heterogeneous newsgroup documents as provided by the AB model. The first column lists the words t which are present in the document n whereas in the second column the most probable aspects k are given along with their posterior probabilities $P(k|t, n, x_{tn})$ ($k \in \{1, ..., 5\}$ in this experiment). Note the uncertainty in explaining some of the more common words.

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 Section 3.3.2, we can correct for the missingness by removing the "phantom" aspect and reconstructing the data again. More precisely, let k^* denote the phantom aspect; this is easily identified by $a_{tk^*} \approx 0 \, \forall t$. Set $s_{k^*n} = 0 \, \forall n$ and normalise all s_{kn} such that $\sum_k s_{kn} = 1$ holds again for all n. Then compute the reconstruction of the data by rounding $p_{tn} = \sum_k a_{tk} s_{kn}$ to 0 or 1, where s_{kn} was updated as described above.

For comparison, we also reconstruct the data by other methods: For 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 identified by any one component but instead the components collaborate in explaining the data as it is.

It should also be noted that models not designed for binary data are somewhat problematic to employ, due to the lack of suitable probabilistic interpretation: With NMF, values

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 algorithm 0.76
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 diet 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 food 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 clinic 0.23
peopl effect doctor medic 0.98 patient 0.87 diseas 0.85 treatment 0.84 medicin 0.84 physician 0.81 food 0.81 symptom 0.80 med 0.80
peopl sin love christ rutger geneva jesu god 0.99 christian 0.99 church 0.79 word 0.79 bibl 0.78 faith 0.78 agre 0.74 accept 0.73 scriptur 0.73
peopl public system agre faith accept christ teach clinic mission religion jesu holi doctrin scriptur god 0.05 christian 0.05 rutger 0.04 word 0.03 church 0.03 bibl 0.03 love 0.03 man 0.03 truth 0.03
govern peopl christ food rutger church atho god 0.74 christian 0.74 word 0.66 accept 0.64 bibl 0.64 faith 0.64 jesu 0.63 agre 0.63 effect 0.63

Table 3: Expansion of 10 randomly selected documents from the 4 Newsgroups collection. For each document, the first line of the cell 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 $p(k|t, n, x_{tn})$ of the phantom.

above 1 are possible as NMF does not treat the values as probabilities, so we simply turn those to 1. In PLSA, the data model is quite different, as the parameters give $p(t|n)$, the probability of generating word t into any word position of document n. 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

$$p(\text{'attribute } t \text{ appears at least once in observation } n\text{'}) = 1 - (1 - p(t|n))^{L_n} \tag{28}$$

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 values when the data is reconstructed as described above, using AB, MB, LPCA, NMF and PLSA. The decrease is largest in post-processed AB; the result for non-post-processed AB is also given for comparison. It is well possible that new missing values 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.

Detecting added noise. A more challenging setting, from the point of demonstrating the abilities of AB, 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"

[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.

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

Table 4: Decrease in the number of missing values when the palaeontological data is reconstructed using the model parameters. Generation of new missing values is possible, as indicated by the negative decrease of PLSA.

or "added". 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 $P(k|t, n, x_{tn}) = q_{k,t,n,x_{tn}}$ that the aspect k has generated the observation x_{tn} is computed as in Formula (4). The histograms of the posteriors $P(k|t, n, x_{tn})$ for true 1s, added 1s and 0s are seen in Figure 8. They are computed as

$$p(k|\text{true ones}) \propto \sum_{t,n:x_{tn}=1 \text{ originally}} P(k|t, n, x_{tn}) \tag{29}$$

$$p(k|\text{added ones}) \propto \sum_{t,n:x_{tn}=1 \text{ added}} P(k|t, n, x_{tn}) \tag{30}$$

$$p(k|\text{zeros}) \propto \sum_{t,n:x_{tn}=0} P(k|t, n, x_{tn}) \tag{31}$$

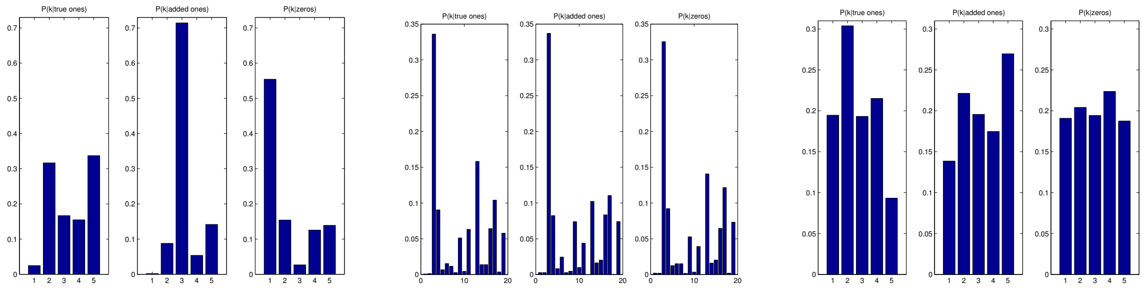
The quantities (29-31) are normalised such that each of them sums to 1 over k. We can see in Figure 8 (a) that the "white phantom" (the leftmost bar in all plots) has a very small or zero probability in true or added 1s and correspondingly a high probability at zeros. The "black phantom" (the third bar 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, Figures 8 (b) and (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. For LPCA and NMF, the quantities (29-31) cannot directly be computed, as the posterior of a component is not a well defined concept.

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

[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.



(a) (b) (c)

Figure 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 (29), the middle one at added 1s (30) and the rightmost at 0s (31). 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.

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.

Out of sample likelihood. The degree of corrosion, measured as the proportion of extra 0s in the data[8], is selected randomly for each observation from an uniform distribution on $[0, 0.4]$. Figure 9 (a) shows the 10-fold cross-validated out of sample likelihood for varying the number of latent components. The behaviour of different models is quite similar to what was seen on the non-corrupted data in Figure 2: LPCA over-fits quickly as the number of latent components increases, and AB always outperforms MB.

We then vary the degree of corrosion so that the proportion of extra 0s in each observation is distributed according to Uniform$[0, u]$ where u is fixed for all observations; separate experiments are conducted on $u = 0.2, 0.4, 0.6, 0.8$. The optimal number K of latent components varies according to u and is selected based on in-sample values of the AIC at each u. The out-of-sample likelihood is then computed and shown in Figure 9 (b) for different choices of u at both AB and MB. We see that as the proportion of extra 0s increases, the out of sample likelihood decreases. From this observation we can draw a new hypothesis regarding another field of application, namely text document data, in which a term absence (0) either means that the term does not fit the topical content of the document, or that the term is simply omitted. It is often argued that a Multinomial model is more powerful than a Bernoulli model for analysing term-document data, as a Multinomial model takes into account the actual frequencies of terms in documents instead of just presences and

[8] Again, this is not the proportion of 1s turned to 0s but indeed the proportion of added 0s in the whole observation. Superimposing an extra 0 where the data instance is 0 does not have any effect.

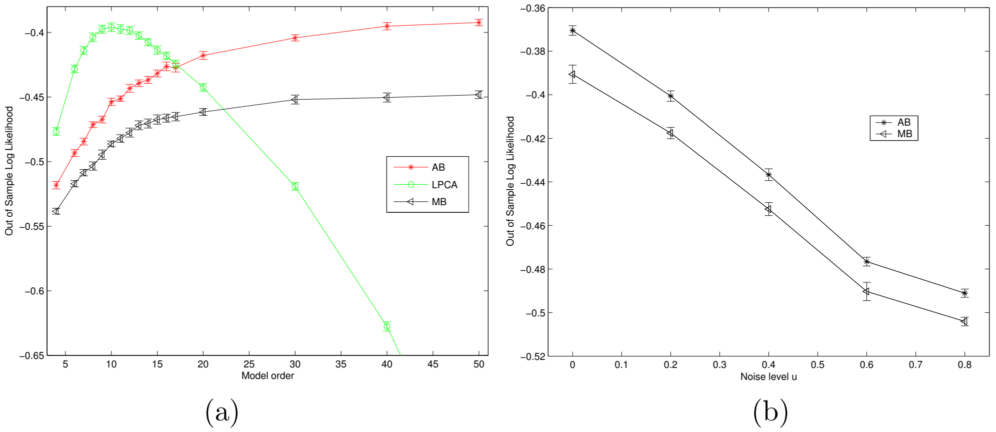


Figure 9: (a) Out of sample log likelihood for the corrupted handwritten digits data set. The error bars show the variation in ten-fold cross-validation (one standard error). Corrosion was added in the data such that the proportion of extra zeros was drawn from a Uniform$[0, 0.4]$ distribution at each observation. (b) Out of sample log likelihoods of the AB and MB models when the proportion of extra 0s is Uniform $[0, u]$. Horizontal axis: u. $u = 0$ refers to the noiseless case.

absences. There may however be another reason, too. A Bernoulli model treats both zeros and ones as information-bearing entities. In text document data, many of the zeros are “missing presences” of terms, in the sense that a term is not used despite it suits well the topical content of the document — in other words, there is a lot of noise in the zeros. A Bernoulli model attempts to model these noisy zeros, too, and risks to fail at generalising to unseen documents.

Noise removal. We then demonstrate the use of the Aspect Bernoulli 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 Section 3.4.1, we identify k^* as the aspect corresponding to noise, by $a_{tk^*} \approx 0 \, \forall t$[9]. 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 10 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 Section 3.4.1 related to Table 4.

In Figure 10 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 again overfits quickly, and PLSA is not very successful: in both methods, the

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

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.

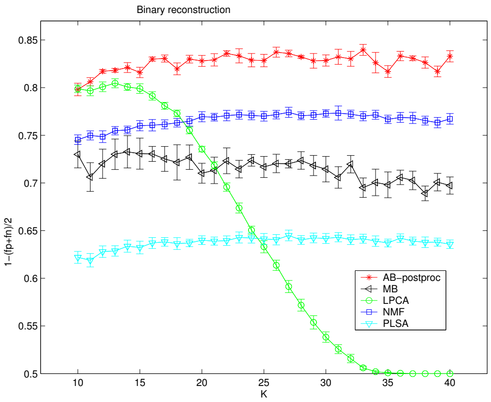


Figure 10: Noise removal in artificially corrupted binary handwritten digit images. The parameters of AB are post-processed by removing the component explaining the noise, re-normalising the parameters, and reconstructing the data. In the other methods, the data is reconstructed from the original estimated parameters, as no single component explains the noise.

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 Figures 9 (a) and 10. The number of latent aspects was chosen based on the AIC as $K = 14$. The top row of Figure 11 shows the 14 bases (parameters a_{tk}) obtained for this data set. In addition to bases that look like prototypical images as they contain high probabilities on corresponding pixels, we also have 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 $P(k|n, t, x_{tn})$ that the k'th 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 Figure 11) 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.

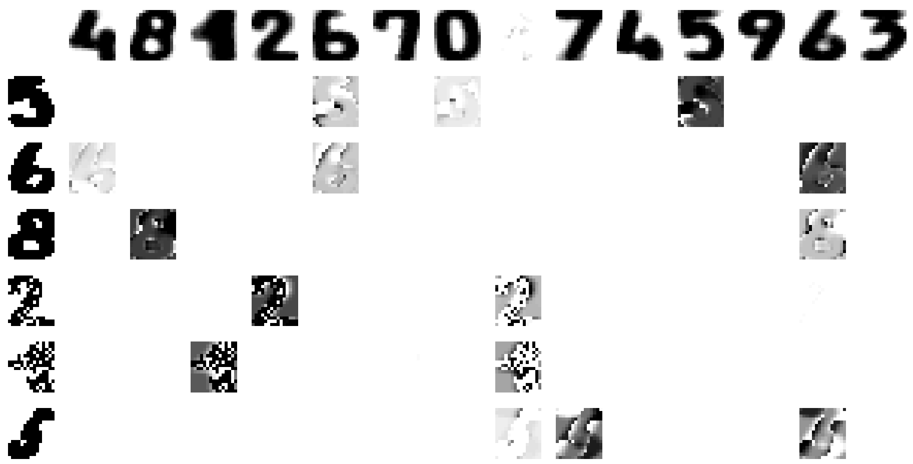


Figure 11: Results on artificially corrupted binary handwritten digit images where some pixels have been turned to white. The images on the top line depict the reshaped parameters a_{tk} as basis images. Some examples from this data set are shown in the first column, and their analysis as provided by the AB model in the next columns. For each datum instance n and each aspect k, the probability values $P(k|n, t, x_{tn})$ are shown for each pixel $t \in \{1, \dots, 240\}$. On all these plots, the level of darkness of a pixel is proportional to the probability of it being 'on'.

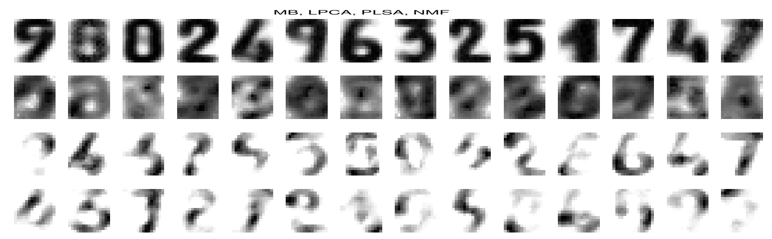


Figure 12: Basis images estimated by MB (top row), LPCA (second row), PLSA (third row) and NMF (bottom row) in artificially corrupted binary handwritten digits.

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 Figure 12. No single basis corresponds to the corruption, instead the bases resemble parts of digits. For the ease of comparison, $K = 14$ bases are estimated; however, the results at different K are quite similar.

Similarly, in Figure 13 the results of AB are shown for the case of added 1s: the proportion of pixels turned "on" was drawn from a Uniform[0,0.4] distribution. There are $K = 15$ bases as chosen by the Akaike Information Criterion; the second to last of them is a “black phantom”. Again, the phantom has a high posterior probability of having created the non-content-bearing black pixels, and the content-bearing pixels (both white and black) are explained by one or a few content-bearing latent aspects.

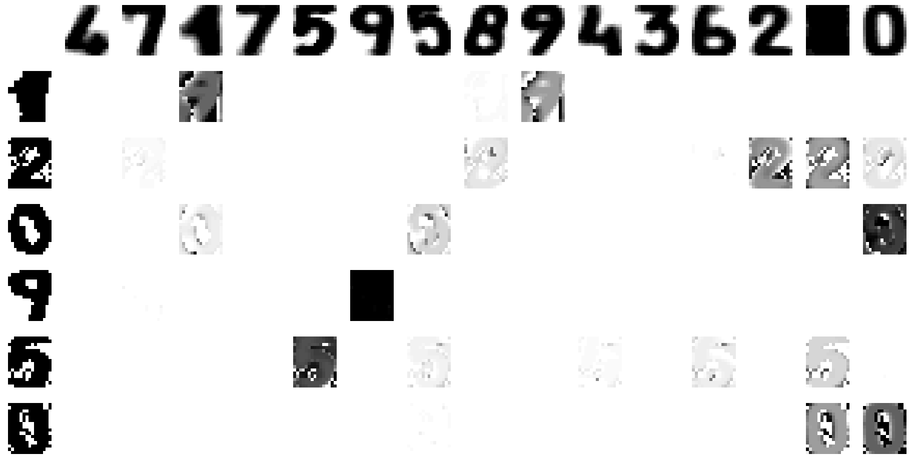


Figure 13: Results on binary handwritten digit images where extra black pixels are added into the images. The images on the top line depict the reshaped parameters a_{tk} as basis images. Some examples from this data set are shown in the first column, and their analysis as provided by the AB model in the next columns. For each datum instance n and each aspect k, the probability values $P(k|n, t, x_{tn})$ 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’.

Again, the bases given by MB, LPCA, PLSA and NMF in the case of added 1s are shown in Figure 14. No single basis corresponds to the added pixels, instead the bases resemble parts of digits very similarly to Figure 12.

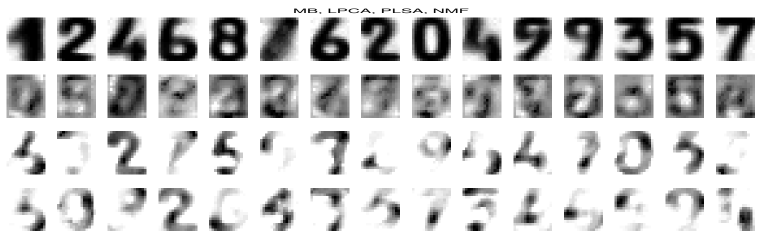


Figure 14: Basis images estimated by MB (top row), LPCA (second row), PLSA (third row) and NMF (bottom row) in binary handwritten digits where extra black pixels are added.

3.4.3 Detecting omitted or added words from Usenet text messages

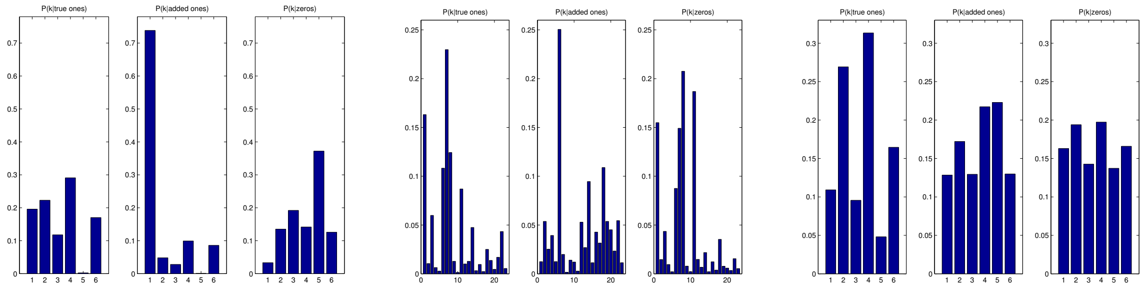
In Section 3.3.3 we have seen that term occurrences in text messages naturally contain a factor of word omission and AB is able to detect that factor. However, the goodness of this detection can only be assessed in a subjective manner. In order to conduct an objective evaluation we would need to have the true values. For this reason, in the same way as with the palaeontological data, we create an artificial setting which we use to objectively demonstrate the ability of the Aspect Bernoulli model in detecting and distinguishing extra, non-content-bearing 1s in binary coded newsgroup documents. 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% — more than doubled. We then estimate $K = 6$ latent aspects in the corrupted data, and obtain four aspects reflecting the four 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.

The latent aspects can also be visualised by listing the keywords having the largest probabilities a_{tk} of being generated by each topic k; these are seen in Table 5 for AB. Again, the term lists of the phantoms do not show any coherence, whereas the real topics are easily distinguished by their keywords.

"black phantom"	*cryptographic*	*medical*	*religious*	*"white phantom"*	*space-related*
33.4	46.9	61.2	51.5	111.1	43.8
effect 1.00	kei 1.00	peopl 0.65	christian 1.00	bless 4.3e-02	space 1.00
symptom 0.99	encrypt 0.95	effect 0.53	god 1.00	man 1.5e-02	nasa 0.99
kei 0.99	govern 0.86	medic 0.45	peopl 0.92	pat 2.6e-03	system 0.92
pgp 0.96	chip 0.85	patient 0.42	rutger 0.81	space 1.3e-07	orbit 0.91
biblic 0.95	public 0.76	doctor 0.34	christ 0.73	den 7.6e-13	cost 0.74
secur 0.93	clipper 0.76	diseas 0.31	bibl 0.70	peopl 8.9e-14	launch 0.68
holi 0.91	system 0.73	med 0.29	church 0.70	agre 8.0e-15	spacecraft 0.68
devic 0.90	secur 0.66	diet 0.28	word 0.65	henri 5.0e-16	mission 0.64
patient 0.89	escrow 0.59	infect 0.28	accept 0.64	satellit 4.9e-16	shuttl 0.63
secret 0.89	comput 0.49	treatment 0.27	agre 0.61	comput 2.9e-16	flight 0.58

Table 5: Six aspects inferred by AB from a document collection of 4 Newsgroup messages where extra terms were randomly added in the documents. Besides four aspects representing the topical features of discussion, there are two "phantom" aspects: the first aspect explains the artificially added 1s and the fifth aspect explains the natural term omissions. For each aspect k, the terms t having the largest generation probabilities a_{tk} are listed, together with the values of the a_{tk}. The top row of each column gives $\sum_k s_{kn}$.

As the document corruption was created artificially by adding extra terms into the documents, we can measure the degree to which the models are able to distinguish between different 1s. Figure 15 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. In Figure 15 (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 Figure 15 (b), 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 Figure 15 (c), none of the PLSA components deviates.



(a) (b) (c)

Figure 15: 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 (29), the middle one at added 1s (30) and the rightmost at 0s (31). In AB, aspects $k = 1$ and 5 differentiate between the three cases.

4. Conclusions

This paper presented a probabilistic multiple cause model for 0-1 data. The Aspect Bernoulli model analyses the causes behind not only the presences (1) but also the absences (0) of attributes, in contrast to all existing models for 0-1 data. A distinctive feature of the Aspect Bernoulli model is its ability to detect 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 as the algorithm minimises the entropies of the model parameters.

In addition to the variety of related factorisation models discussed, it might also be interesting to follow up some similarity with the model of Law et al. (2004), proposed for feature selection in unsupervised learning, in a mixture of Gaussians setting. The data generation process in their model is remotely similar to our AB model: For each observation, a content-bearing latent cause k is first selected. 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 “common cause” can be seen in analogy with our phantom aspect. However, in contrast to AB, only one content-bearing cause is used for all attributes in an observation, and the model is thus not a multiple cause model in a strict sense.

In a somewhat similar line of thought, the models of Hofmann (1999), Barnard et al. (2001, 2003) and Blei et al. (2004) present hierarchical architectures where the latent components are arranged into a tree, and a multidimensional observation is constructed by travelling down the tree and generating multinomial attribute occurrences according to the latent components visited on the path of the tree. The root node is a common component that may participate in the generation of all the observations. Again a remote analogy can be followed in that this common component often captures uninformative features such as the stop words in documents. Recently, using a somewhat similar tree-construction has been considered explicitly for finding uninformative features by Wang and Kabán (2005). However, our factorisation approach separates out the common (non-specific, un-informative) features without imposing any artificial topologies.

As a further refinement of the Aspect Bernoulli model one could consider a fully symmetric model that analyses the observations and attributes simultaneously: in some data sets such as the palaeontological data considered in this paper, there are both observation-dependent and attribute-dependent noise factors, requesting a model that is capable of considering both dimensions.

An intermediate model between Logistic PCA and Aspect Bernoulli could also be constructed for completeness, as $p(\mathbf{x}|\mathbf{s},\mathbf{a}) = \prod_n \prod_t g(\sum_k a_{tk} s_{kn})^{x_{tn}} (1 - g(\sum_k a_{tk} s_{kn}))^{1-x_{tn}}$ 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 Aspect Bernoulli 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.

We have also demonstrated how the Aspect Bernoulli 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. Non-Bernoulli models such as probabilistic latent semantic analysis and nonnegative matrix factorisation cannot be compared to in terms of out-of sample likelihoods, as the event spaces are different; also, they cannot separate binary noise into one single component, making noise removal more difficult.

Acknowledgements

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

Appendix A

The derivation of the algorithm in (9-11) is as follows.

Denote $\overline{a_{tk}} = 1 - a_{tk}$. The log likelihood (8) 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:

$$\begin{aligned}\mathcal{L} = \sum_n \sum_t [x_{tn} \log \sum_k a_{tk} s_{kn} + (1 - x_{tn}) \log \sum_k \overline{a_{tk}} s_{kn} \\ - c_{tk}(a_{tk} + \overline{a_{tk}} - 1) - \lambda_n (\sum_k s_{kn} - 1)]\end{aligned} \tag{32}$$

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

$$\frac{\partial \mathcal{L}}{\partial a_{tk}} = \sum_n \frac{x_{tn}}{\sum_\ell a_{t\ell} s_{\ell n}} s_{kn} - c_{tk} = 0 \tag{33}$$

$$\frac{\partial \mathcal{L}}{\partial \overline{a_{tk}}} = \sum_n \frac{1 - x_{tn}}{\sum_\ell \overline{a_{t\ell}} s_{\ell n}} s_{kn} - c_{tk} = 0 \tag{34}$$

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

$$a_{tk} \sum_n \frac{x_{tn}}{\sum_\ell a_{t\ell} s_{\ell n}} s_{kn} - c_{tk} a_{tk} = 0 \tag{35}$$

$$\overline{a_{tk}} \sum_n \frac{1 - x_{tn}}{\sum_\ell \overline{a_{t\ell}} s_{\ell n}} s_{kn} - c_{tk} \overline{a_{tk}} = 0 \tag{36}$$

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

$$c_{tk} = a_{tk} \sum_n \frac{x_{tn}}{\sum_\ell a_{t\ell} s_{\ell n}} s_{kn} + \overline{a_{tk}} \sum_n \frac{1 - x_{tn}}{\sum_\ell \overline{a_{t\ell}} s_{\ell n}} s_{kn} \tag{37}$$

as in (11). From (35) we have the solution for a_{tk} in the form of a fixed point equation:

$$a_{tk} = a_{tk} \sum_n \frac{x_{tn}}{\sum_\ell a_{t\ell} s_{\ell n}} s_{kn} / c_{tk} \tag{38}$$

as in (10). Solving for s_{kn} proceeds similarly: the stationary equation is

$$\frac{\partial \mathcal{L}}{\partial s_{kn}} = \sum_t \left(\frac{x_{tn}}{\sum_\ell a_{t\ell} s_{\ell n}} a_{tk} + \frac{1 - x_{tn}}{\sum_\ell \overline{a_{t\ell}} s_{\ell n}} \overline{a_{tk}} \right) - \lambda_n = 0 \tag{39}$$

Multiplying both sides by s_{kn} we obtain

$$\sum_t \left(\frac{x_{tn}}{\sum_\ell a_{t\ell} s_{\ell n}} a_{tk} s_{kn} + \frac{1 - x_{tn}}{\sum_\ell \overline{a_{t\ell}} s_{\ell n}} \overline{a_{tk}} s_{kn} \right) = \lambda_n s_{kn} \tag{40}$$

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

$$\lambda_n = \sum_t \frac{x_{tn}}{\sum_\ell a_{t\ell} s_{\ell n}} \sum_k a_{tk} s_{kn} + \sum_t \frac{1 - x_{tn}}{\sum_\ell \overline{a_{t\ell}} s_{\ell n}} \sum_k \overline{a_{tk}} s_{kn} = \sum_t x_{tn} + \sum_t (1 - x_{tn}) = T \tag{41}$$

Having computed λ_n, from (40) we obtain the fixed point equation for s_{kn}, identical to (9):

$$s_{kn} = s_{kn} \{ \sum_t \frac{x_{tn}}{\sum_\ell a_{t\ell} s_{\ell n}} a_{tk} + \frac{1 - x_{tn}}{\sum_\ell \overline{a_{t\ell}} s_{\ell n}} \overline{a_{tk}} \} / T \tag{42}$$

Here we presented a derivation of (9)-(11) which is independent of the model formulation provided at (1). Naturally, the same equations could be obtained by starting from the algorithm (4)-(6): replacing (4) into (5) and (6), and manipulating the resulting expressions, using that x_{tn} is either 0 or 1.

References

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast discovery of association rules. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.), *Advances in Knowledge Discovery and Data Mining*, chap. 12, pp. 307–328. AAAI Press.

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.

Barlow, H., Kaushal, T., & Mitchison, G. (1989). Finding minimum entropy codes. *Neural Computation*, pp. 412–423.

Barnard, K., Duygulu, P., & Forsyth, D. (2001). Clustering art. In *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. II:434–441.

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., & Jordan, M. I. (2003). Matching words and pictures. *Journal of Machine Learning Research*, *3*, 1107–1135.

Barry, J. C., 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)*, 1–71.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian Theory.* Wiley.

Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. In Thrun, S., Saul, L., & Schölkopf, B. (Eds.), *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Buntine, W. (2002). Variational extensions to EM and multinomial PCA. In *Machine Learning: ECML 2002*, No. 2430 in Lecture Notes in Artificial Intelligence (LNAI), pp. 23–34. Springer-Verlag.

Collins, M., Dasgupta, S., & Schapire, R. E. (2001). A generalization of principal component analysis to the exponential family. In *Advances in Neural Information Processing Systems 14*, pp. 617–624.

Dayan, P., & Zemel, R. S. (1995). Competition and multiple cause models. *Neural Computation*, *7*(3), 565–579.

Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. Tech. rep. TR 2001-05, Department of Computer Sciences, University of Texas, Austin.

Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions*. Chapman & Hall, London.

Földiák, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*, *64*, 165–170.

Fortelius, M., Werdelin, L., Andrews, P., Bernor, R. L., 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*, pp. 414–448. Columbia University Press.

Gordon, G. J. (2003). Generalized^2 linear^2 models. In Becker, S., Thrun, S., & Obermayer, K. (Eds.), *Advances in Neural Information Processing Systems 15*, pp. 577–584. MIT Press, Cambridge, MA.

Gyllenberg, M., Koski, T., Reilink, E., & Verlaan, M. (1994). Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, *31*, 542–548.

Harman, H. H. (1967). *Modern Factor Analysis* (2nd edition). University of Chicago Press.

Heckerman, D., & Chickering, D. M. (1996). A comparison of scientific and engineering criteria for Bayesian model selection. Tech. rep., Microsoft Research.

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*, pp. 682–687. Morgan Kaufmann.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*, 177–196.

Hofmann, T. (2004). Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, *22*(1), 89–115.

Hofmann, T., & Puzicha, J. (1998). Unsupervised learning from dyadic data. Tech. rep. TR-98-042, Berkeley, CA.

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*, pp. 688–693. Morgan Kaufmann.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent Component Analysis.* Wiley Interscience.

Jaakkola, T. S. (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.

Jolliffe, I. T. (1986). *Principal Component Analysis.* Springer.

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.

Kemp, C., Griffiths, T. L., & Tenenbaum, J. B. (2004). Discovering latent classes in relational data. Tech. rep. AI Memo 2004-019, Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory.

Law, M. H. C., Figueiredo, M. A. T., & Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(9), 1154–1166.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788–791.

Lee, D. D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*, pp. 556–562.

Marlin, B. (2004). Modeling user rating profiles for collaborative filtering. In Thrun, S., Saul, L., & Schölkopf, B. (Eds.), *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA.

Marlin, B., & Zemel, R. (2004). The multiple multiplicative factor model for collaborative filtering. In *ICML-2004: Proceedings of the 21st International Conference on Machine Learning*, pp. 576–583.

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, pp. 41–48. AAAI Press.

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*, pp. 249–257. Morgan Kaufmann Publishers Inc.

Polčicová, G., & Tiňo, P. (2004). Making sense of sparse rating data in collaborative filtering via topographic organization of user preference patterns. *Neural Networks*, *17*, 1183–1199.

Puolamäki, K., Fortelius, M., & Mannila, H. (2005). Seriation in paleontological data using Markov Chain Monte Carlo methods. Submitted.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press.

Saund, E. (1995). A multiple cause mixture model for unsupervised learning. *Neural Computation*, *7*(1), 51–71.

Schein, A., Saul, L., & Ungar, L. (2003). A generalized linear model for principal component analysis of binary data. In Bishop, C. M., & Frey, B. J. (Eds.), *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics.*

Schmidhuber, J. (1992). Learning factorial codes by predictability minimization. *Neural Computation*, *4*(6), 863–879.

Seppänen, J. K., 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*, No. 2838 in Lecture Notes in Artificial Intelligence, pp. 423–434. Springer.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In Rumelhart, D. E., & McClelland, J. L. (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1: Foundations, pp. 194–281. MIT Press.

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validatedlikelihood. *Statistics and Computing*, *10*(1), 63–72.

Srebro, N. (2004). *Learning with Matrix Factorizations.* Ph.D. thesis, Massachusetts Institute of Technology.

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)*, pp. 720–727. AAAI Press.

Tipping, M. E. (1999). Probabilistic visualisation of high-dimensional binary data. In Kearns, M. S., Solla, S. A., & Cohn, D. A. (Eds.), *Advances in Neural Information Processing Systems 11*, pp. 592–598.

Wang, X., & Kabán, A. (2005). Finding uninformative features in binary data. In Gallagher, M., Hogan, J. M., & Maire, F. (Eds.), *Intelligent Data Engineering and Automated Learning - IDEAL 2005*, Vol. 3578 of *Lecture Notes in Computer Science*, pp. 40–47. Springer.

Zemel, R. S. (1993). *A minimum description length framework for unsupervised learning.* Ph.D. thesis, University of Toronto.