Learning to Read Between the Lines: The Aspect Bernoulli Model

A. Kabán[*] E. Bingham[†] T. Hirsimäki[†]

Abstract

We present a novel probabilistic multiple cause model for binary observations. In contrast to other approaches, the model is linear and it infers reasons behind both observed and unobserved attributes with the aid of an explanatory variable. We exploit this distinctive feature of the method to automatically distinguish between attributes that are 'off' by content and those that are missing. Results on artificially corrupted binary images as well as the expansion of short text documents are given by way of demonstration.

1 Introduction

Developing generative models for inferring multiple reasons behind observations has long been a central aim of unsupervised learning [10, 4, 2]. Besides variations determined by the type of data at hand, models vary in the form of interactions they assume the hidden generators may follow in explaining (generating) the observations.

In this paper we will concentrate on modelling binary coded data where only the presence or absence of an attribute is of interest; this is in contrast to count data in which the actual frequencies of attributes are coded. For count data, multinomial models [4, 2, 3] are a common choice whereas for binary data, Bernoulli mixtures [6, 7] among others have been used. It is worth noticing that while multinomial models [4, 2] are focusing exclusively on what has been explicitly observed, Bernoulli models by definition need to explain both the presence and the absence of the attributes, that is, both the zeros and ones in the data. In other words, in a multinomial model, the statistical events are the attributes, whereas in Bernoulli models the statistical events are 'on' (1) or 'off' (0) [8]. However, in existing approaches to multiple-cause modelling of binary observations [10, 11], the attributes that are 'off' are suppressed, and the effort is spent on identifying the difficult nonlinear relationship between hidden causes and the attributes that are 'on'. Indeed, this relationship may take the form of a discrete logical OR in which case NP-completeness of the problem has been proven [11], or a noisy-OR [10] in which case a gradient-based optimisation is provided in [10] as no closed form solution is available.

However, while this nonlinear modelling may be essential in some applications, there are also cases, when inferring reasons behind having not observed some of the attributes is at least as important as those behind having observed some others. Examples include images with a varying degree of corrosion, or text corpora where words might be absent either because their presence would be in contradiction with the topical content of a document, or simply because they are missing.

In this paper we formulate a linear multiple-cause model for multivariate binary data which yields a novel method of dealing with this problem in a principled manner. The model can be formally seen as a Bernoulli analogue of the multinomial decomposition model known under the names of aspect model, probabilistic Latent Semantic Analysis [4] and also as multinomial Principal Component Analysis [3]. For placing the proposed model in context, we will show that — as opposed to logistic distributed models [13, 14], which are nonlinear and proceed by decomposing the Bernoulli natural parameter, our model performs a convex decomposition of the mean parameter of the Bernoulli. Discussions and illustrative examples will be provided.

2 The aspect Bernoulli model

Let $\boldsymbol{x}_n$ denote a T-dimensional multivariate binary observation and x_{tn} its t-th component, $t = 1, \ldots, T$. A multiple-cause generation process on the level of the multivariate observations allows the components of the instance $\boldsymbol{x}_n$ to be generated by different latent causes $k = 1, \ldots, K$ in instance-specific proportions. The n-th instance is assumed to be generated as the following:

- Pick a discrete distribution $P(.|n)$ over the latent causes from the set of all K-dimensional discrete distributions (uniformly)
- For each component x_{tn} of $\boldsymbol{x}_n$, $t = 1, \ldots, T$:
  Pick a cause k with probability $P(k|n)$.
  Pick either an 'on' (1) or an 'off' (0) event for

[*] Dept. of Computer Science, University of Birmingham, B15 2TT, UK. Email A.Kaban@cs.bham.ac.uk. The work has been performed while visiting Helsinki University of Technology.

[†] Lab. of Computer and Information Science, Helsinki University of Technology, Finland. Email ella@iki.fi, teemu.hirsimaki@hut.fi

x_{tn} according to a component-specific univariate Bernoulli distribution, conditioned on k.

Denoting the instance-specific probability of the k-th aspect by $s_{kn} = P(k|n)$ and the conditional probability of the t-th component being 'on' by $P(1|k,t) = a_{tk}$, the conditional data likelihood of this model is the following.
(2.1)

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

where $\boldsymbol{s}_n$ is a vector of the probabilities s_{kn}. We made the modelling assumption here that data components are conditionally independent given the latent variable, to force their dependencies to be captured by the lower dimensional hidden variable $\boldsymbol{s}$. Assuming a uniform Dirichlet prior on $\boldsymbol{s}$, then $p(\boldsymbol{x}_n|\boldsymbol{s}_n) \propto p(\boldsymbol{x}_n, \boldsymbol{s}_n) \propto p(\boldsymbol{s}_n|\boldsymbol{x}_n)$. Maximising this quantity will therefore yield a maximum a posteriori (MAP) / Maximum Likelihood (ML) estimation algorithm as derived below. This provides the most probable hypothesis $\boldsymbol{s}$, which is the optimal choice when the main focus is to study the representational characteristics of a model on fixed data sets [9] — which is the primary scope in this paper. Approximate estimates of a fully generative model formulation similar to the approach taken in [2] can also be straightforwardly obtained, however this is outside the scope of this paper.

Taking an EM-type approach [1, 5] to maximising (2.1), we obtain the iterative algorithm below.

$$q_{k,n,t,x_{tn}} \propto s_{kn} a_{tk}^{x_{tn}} (1-a_{tk})^{1-x_{tn}} \tag{2.2}$$

$$s_{kn} = \sum_t q_{k,n,t,x_{tn}}/T \tag{2.3}$$

$$a_{tk} = \frac{\sum_n x_{tn} q_{k,n,t,x_{tn}}}{\sum_n q_{k,n,t,x_{tn}}} \tag{2.4}$$

Note that the constraint $a_{tk} \in [0,1]$ needs not be explicitly imposed in this model setting, as it will be automatically satisfied given that the other constraints are satisfied and the data is binary — as can be seen from the form of (2.4)[1]. Here $q_{k,n,t,x_{tn}}$ represents the posterior probability that the cause k has generated the observation (either the 0 or the 1) at the t-th component of the n-th instance, i.e. $P(k|n,t,x_{tn})$.

A somewhat similar procedure has been recently developed for the case of continuous data in [5], based on Gaussians and utilised for analysing / predicting user ratings.

2.1 An alternative view It is also interesting, from the point of relating this model to other binary models, to point out that (2.1) can be rewritten as the following.
(2.5)

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

To see this, notice that when $x_{tn} = 1$, then according to both (2.1) and (2.5) we have that $p(x_{tn}|\boldsymbol{s}_n) = \sum_k a_{tk} s_{kn}$; and for the case when $x_{tn} = 0$ we have $p(x_{tn}|\boldsymbol{s}_n) = \sum_k (1 - a_{tk}) s_{kn}$ from both (2.1) and (2.5). In obtaining the latter equality, we have used the convexity of the combination — indeed, note that $1 - \sum_k a_{tk} s_{kn} = \sum_k (1 - a_{tk}) s_{kn}$.

We can now observe that, similarly to the multinomial aspect model, the mean parameter of the distribution (Bernoulli in this case), for the t-th component of the n-th observation, is factorised in a convex combination: $p_{tn} := p(x_{tn} = 1|\boldsymbol{s}_n) = \sum_k a_{tk} s_{kn}$. Although the generative multinomial aspect model is also known under the name of multinomial PCA in the literature [3], the term Bernoulli PCA would be somewhat confusing, as the extension of the PCA technique for Bernoulli data [13] refers to logistic PCA which factorises the natural parameter of the Bernoulli: $\theta_{tn} := \sum_k a_{tk} s_{kn}$, where $p_{tn} = e^{\theta_{tn}}/(1 + e^{\theta_{tn}})$. Therefore the model just described should rather be termed as aspect Bernoulli (AB) model. In terms of the restrictions imposed on the model parameters, the AB model lies between logistic PCA [13] (which does not restrict the values of a_{tk} and s_{kn}, in contrast to AB) and single-cause Bernoulli mixtures [7] (which assume one hidden cause for a multivariate observation). Implications of this regarding various criteria such as compression accuracy, generalisation and direct interpretability of the parameters will be demonstrated in the next section.

3 Simulations

3.1 Compression and generalisation on binary digit images The data set utilised for the first demonstration is a collection of 1,000 binary digital images containing handwritten digits[2]. There are 200 instances from each digit category, each image containing 15×16 pixels, each of which can be either 'on' or 'off'.

We demonstrate the Aspect Bernoulli (AB) model in the context of two related Bernoulli models, Logistic Principal Component Analysis (LPCA) [13] and single cause mixture of Bernoullis (MB). As already pointed out, AB is more restrictive than LPCA but more general than MB. As expected, in terms of compression (reconstruction) of a fixed data set, measured as the

[1] This is a consequence of the fact that the Bernoulli is a member of the exponential family of distributions and follows from the first moment identity.

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

negative log likelihood, AB lies between LPCA and MB, as can be seen on the left hand plot of Figure 1. The best results over 15 randomly initialised restarts have been retained for all models in this experiment, in order to avoid problems of local optima due to the non-convex optimisation.

However, in terms of generalisation the picture looks different on this data. We have evenly divided the data set into a training set and an independent test set and have employed an empirical Bayesian latent density estimate in computing out of sample likelihood values:

$$-\log \frac{1}{N} \sum_{n} \prod_{t} (\boldsymbol{a}_t \boldsymbol{s}_n)^{x_{t,test}} (1 - \boldsymbol{a}_t \boldsymbol{s}_n)^{1 - x_{t,test}}$$

and likewise for LPCA. Here $\boldsymbol{a}_t$ denotes the t-th row of $\boldsymbol{A}$, $\boldsymbol{s}_n$ is the latent vector obtained for the training point $\boldsymbol{x}_n$ and $x_{t,test}$ is the t-th dimension of a new, previously unseen test point (image). Results are shown on the right hand plot of Figure 1. The LPCA experiences overfitting after $K = 7$ aspects (7 assumed causes) whereas AB only starts to overfit after $K = 40$ aspects. The best generalisation performance of the MB is inferior to those of both LPCA and AB on this data set.

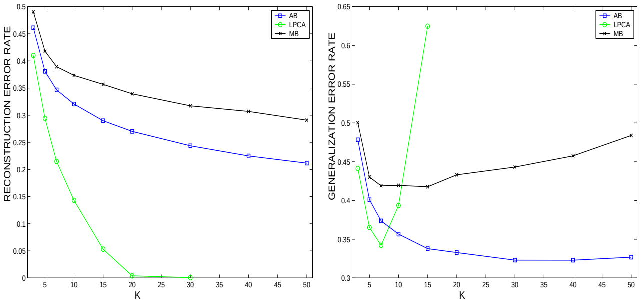

Figure 1: Reconstruction error (left) and generalisation error (right) obtained on the handwritten digits data set. Horizontal axis: K.

3.2 Multiple cause explanatory power In order to demonstrate the distributive nature of the AB representation and the interpretability of the AB parameters, it is more interesting to see the aspect Bernoulli model at work when there may be several causes for the same observation, e.g. when corrosion is present in the data. In its initial form, any pixel that is 'off' in the digits data set can be explained by the category content of the image. However, in a second round of experiments, we have artificially introduced a uniformly varying level of corrosion, by turning off some of the pixels that were 'on' initially on the image. Clearly, by doing so we have created distinct causes for observing the value 'off' for a pixel and identifying this will serve as an evaluation criteria for the proposed multiple-cause model. An AB model with 10 components and the bases obtained are shown in the upper row of Figure 2. It can be observed that in addition to bases that contain high probabilities on bunches of pixels that together look like prototypical images of digits, the AB model has also identified 'phantom-bases' as additional common causes. The first column of Figure 2 shows data instances whose analysis is provided in the remainder of the columns. For each image, the probability that a basis k explains the observed value (either 0 or 1) of any of its pixels, i.e. $P(k|n, t, x_{tn})$ are shown in column k. On all these plots, the level of darkness of a pixel is proportional to the probability of it being 'on'. For the '1' depicted on the first data instance (second row of Figure 2) we can observe that those pixels which are off due to the artificially created corrosion are explained by one of the 'phantom-bases' with the highest probability. At the same time, the pixels in the upper left and lower right corners of the image, which are 'off' due to the content of the image, are explained by the fourth basis, a '1' that is indeed quite similar to the observed image. Further examples are given on the rows that follow, the last three showing more complex cases where pixels that are 'on' may also have multiple causes. On Figure 3 we show for comparison the basis set (with the same number of components) created by mixture Bernoulli, logistic PCA [13] and NMF [12] models on the same corrupted data. None of these models are able to 'explain' the corrosion as a common cause in an intuitively meaningful way. As in this example the corrosion has been created artificially, then we can objectively measure the degree to which the model is able to distinguish between different reasons. This is shown for the previous experiment in the form of normalised histograms on Figure 4. For each aspect k, the following quantities have been computed:

$$P(k|\text{missing pixels}) \propto \sum_{n,t:x_{tn}=0 \text{ missing}} P(k|n, t, x_{tn})$$

$$P(k|\text{'on'}) \propto \sum_{n,t:x_{tn}=1} P(k|n, t, x_{tn})$$

$$P(k|\text{'off' by content}) \propto \sum_{n,t:x_{tn}=0 \text{ by content}} P(k|n, t, x_{tn})$$

$$P(k|\text{content bearing pixels}) = 1 - P(k|\text{missing pixels}) \\ \propto P(k|\text{'on'}) + P(k|\text{'off' by content})$$

Summarising, from the first histogram we have that 70% of the zeros which are due to corrosion are explained by the 'phantom'-like bases. In contrast, summing the numbers on the third histogram we find that 88% of the content bearing zeros are explained by the 8 content-bearing bases in this experiment. To give more exact numbers in addition to the relative values above, there

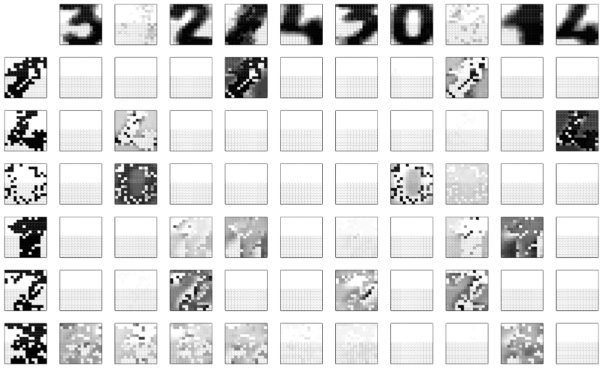

Figure 2: Results on artificially corrupted binary handwritten digit images. The images on the top line depict the reshaped parameters (basis-images) estimated by the aspect Bernoulli model. Some examples from this data set are shown in the first column along with their analysis as provided by the proposed model in the next columns. For each datum instance n and each basis k, the probability values $P(k|n,t,k)$ are shown for each pixel $t \in \{1,...,240\}$. On all these plots, the level of darkness of a pixel is proportional to the probability of it being 'on'.

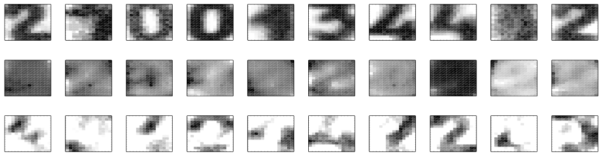

Figure 3: Representation bases created on artificially corrupted binary handwritten digit images by MB (top row), LPCA (middle row) and NMF (last row) respectively. None of these models produce a meaningful distinction from the initial non-corrupted data-set.

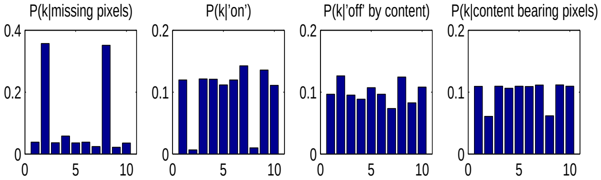

Figure 4: Numerical evaluation of the multiple cause representation provided by the AB model on artificially corrupted digit images.

were a number 18,095 pixels turned off out of 72,677 pixels that were 'on' initially in the non-corrupted data in this experiment.

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

Table 1: Five causes inferred from a document collection from 4 Newsgroup messages.

3.3 Reading between the lines from binary coded Newsgroup messages A real world example that has a similar structure to the one just presented is text. Binary coded text-based data inherently contains missing words — not all words that may express a topic are covered in a document about that topic. However, 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. To illustrate this, we performed the analysis of a subset from the 20Newsgroups collection[3], which contains short Usenet messages from 4 different topics of discussion (4 different newsgroups). A number of 100 documents from each newsgroup were selected. A binary term by document matrix was created using the Bow toolkit[4]. Table 1 provides a typical example of how the aspect Bernoulli model can capture that missing words are a common feature of text. The list of the top few words supported by each of the identified causes, along with their probability of being generated by that cause is summarized (Table 1) for a run with five aspects for a collection of Usenet messages from the following four newsgroups: 'sci.crypt', 'sci.med', 'sci.space' and 'soc.religion.christian'. We have chosen these numbers for the ease of presentation. As we can observe from the table, the second cause is a 'phantom-topic', i.e. an aspect in which being 'on' has a negligible probability for all words (obviously this aspect is responsible for the presence of some of the zeros in the data) — whereas the other four are clearly related to the various topics of discussion. Apart from this distinctive feature, the model is also able to represent

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

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

polysemy – e.g. the word ‘system’ is generated by both the ‘space-related’ and ‘cryptographic’ aspects. The identifiers attached to each cause, shown in the table header, have intentionally been chosen as adjectives, in order to emphasize that these lists represent features that are common to possibly overlapping subsets of the data.

Finally, we show how we can use the proposed method to ‘read between the lines’, i.e. to infer a list of missing words and so expand short text messages. Table 2 provides the top list of the most probable words for which $P(\text{‘}phantom\text{’}|n,t,x_{tn})$ is highest for eight randomly selected documents of the corpus. As expected, these are all words which are not present in the document under consideration, however their absence is not explained by topical features as they are semantically strongly related to the words which are present in the document. Indeed, the document could readily be expanded with the list of words obtained. Investigating the presented model for multiple Bernoulli query expansion will therefore be an interesting future work to pursue.

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

Table 2: Expansion of eight randomly selected documents from the 4 Newsgroups collection. For each document, the first line of the cell contains the words present in the document, followed by the top list of words that the ’phantom-topic’ is responsible for, along with the posterior probability of the ’phantom’ given a document and a word.

4 Conclusions

We have presented a novel probabilistic multiple cause model for inferring hidden causes behind multivariate binary observations. As opposed to the multinomial analogue of the model [4] as well as to previous nonlinear multiple cause models of binary data — some of which try to compensate for a binary thresholding of frequency counts data [10, 11] — the presented model, by its construction, infers reasons behind both the observed and the unobserved attributes and we have exploited this for automatically distinguishing between attributes which are off and those that are missing. Illustrative examples on artificially corrupted digit images as well as binary coded text have been presented and discussed, and comparisons have been shown.

Acknowledgements

The authors wish to thank professor Heikki Mannila for valuable comments on the manuscript.

References

[1] C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.

[2] D.M. Blei, A.Y.Ng and M.I. Jordan, Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(5):993–1022, 2003.

[3] W. Buntine, Variational extensions to EM and multinomial PCA. *Proc. European Conference of Machine Learning*, LNAI 2430, pp.23–34, 2002.

[4] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[5] T. Hofmann, Gaussian latent semantic models for Collaborative Filtering, Proc. ACM SIGIR, 2003.

[6] B. S. Everitt and D. J. Hand, *Finite mixture distributions*, Chapman & Hall, 1981.

[7] M. Gyllenberg, T. Koski, E. Reilink and M. Verlaan, Nonuniqueness in probabilistic numerical identification of bacteria, *J. of Applied Probability*, 31:542–548, 1994.

[8] A. McCallum and K. Nigam, A comparison of event models for Naive Bayes text classification. Proc. of AAAI/ICML-98 Workshop on Learning for Text Categorization, 1998, pp. 41–48.

[9] T. Mitchell, Machine Learning. McGraw-Hill, New York, US, 1996.

[10] E. Saund, A multiple cause model for unsupervised learning. Neural Computation 7:51–71.

[11] J. Seppänen, E. Bingham and H. Mannila, A simple algorithm for topic identification in 0-1 data, Proc. PKDD 2003, pp.423–434.

[12] D.D. Lee and H.S. Seung, Algorithms for non-negative matrix factorization, Advances in Neural Information Processing Systems, 2000, pp.556–562.

[13] A.I. Schein, L.K. Saul, L.H. Ungar, A generalised linear model for Principal Component Analysis of binary data, Proc. 9th International Workshop on Artificial Intelligence and Statistics, January 2003.

[14] M.E. Tipping, Probabilistic visualisation of high dimensional data, Advances in Neural Information Processing Systems, 1999, pp. 592–598.