Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 71 (2008) 2291–2308
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Factorisation and denoising of 0–1 data: A variational approach

Ata Kabán[a,*], Ella Bingham[b]

[a]School of Computer Science, The University of Birmingham, Birmingham B15 2TT, UK
[b]Helsinki Institute for Information Technology, Basic Research Unit, University of Helsinki, P.O. Box 68, Finland

Available online 25 February 2008

Abstract

Presence–absence (0–1) observations are special in that often the absence of evidence is not evidence of absence. Here we develop an independent factor model, which has the unique capability to isolate the former as an independent discrete binary noise factor. This representation then forms the basis of inferring missed presences by means of denoising. This is achieved in a probabilistic formalism, employing independent beta latent source densities and a Bernoulli data likelihood model. Variational approximations are employed to make the inferences tractable. We relate our model to existing models of 0–1 data, demonstrating its advantages for the problem considered, and we present applications in several problem domains, including social network analysis and DNA fingerprint analysis.

Keywords: Factor models; Data denoising; 0–1 data

1. Introduction

Binary data repositories arise from areas as diverse as social sciences, bioinformatics, or forensics research. The processing of binary data requires appropriate tools and methods for tasks such as exploratory analysis, feature construction and denoising. These necessarily must follow the specific distributional characteristics of the data and cannot be accomplished with tools that exist for continuous-valued data analysis.

In particular, in binary data, a '1' encodes the presence, whereas a '0' the absence of an evidence. It is common sense, however, that more often than not, the absence of evidence is not evidence of absence [23]. For example, the pixels of corrupted black and white images, the usage of words in natural language, the presence–absence patterns of social relationships or the entries of a matrix of detections of any kind all typically share this characteristic. In other binary data sets in turn, the absence of evidence is also an evidence of absence—e.g. in clean b&w raster images, the pixels that are present and those that are absent on the image, together define the content of the image. (i) How can we find out whether a given 0–1 data set has such anomalies? (ii) How can we restore a likely 'original'? Currently there is no automated method available to answer these questions, and this is what we tackle in this paper.

We regard (i) as a source separation problem: Besides content-bearing independent factors, we also need to isolate an independent factor that represents absence of evidence but not evidence of absence. If successful, this representation forms a basis for approaching the second part of the problem, (ii), which is essentially a data denoising problem. Note, the order is important here, since the existence of noise is not easily detectable, as the noisy observations are still discrete binary.

Previous successes of factor models and in particular independent component analysis (ICA) [12] make it an important statistical principle worthy of investigation for tackling both explanatory analysis and denoising problems. However, the ICA literature has been developed for continuous-valued observation signals by large, and the particular questions outlined above have never been addressed in the context of 0–1 data. Work on ICA methods for *binary observations* has been very scarce [11,6] despite their wide potential applicability, and related

*Corresponding author.
E-mail addresses: A.Kaban@cs.bham.ac.uk, axk@cs.bham.ac.uk (A. Kabán), ella@iki.fi (E. Bingham).

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.07.038

methods for discrete data in general and binary data in particular are mostly developed outside the 'mainstream' ICA community [25].

Several authors have considered the case of binary sources in the ICA literature, most recently e.g. [8,19] who give algorithms for the under-determined case of less sensors than sources. There are two major differences from this setting though, which make these methods inappropriate for the problem we consider here: First, the unknown components are binary but the noisy observations are real-valued due to the Gaussian noise assumed. As the authors point out, it is then an easy matter to determine whether there is noise or not in the data. By contrary, our observations are always binary. Hence our algorithm needs to be successful in separating out the noise component in order to reveal its presence. This is exactly the problem that we tackle. The noise component is obviously non-Gaussian, still, we will see from the presented applications that it is a very frequently occurring type of noise in real-world 0–1 data. Yet, it was never explicitly noticed in the 0–1 data analysis literature. Secondly, our setting is not under-determined but over-determined. The number of sensors in our case corresponds to the number of samples collected (e.g. number of images, number of text documents, number of nodes in a graph etc.). Although the sample size may be small, it is assumed that the number of components is smaller. In addition, contrary to methods that seek discrete binary sources, in this work, the sources will be allowed to take continuous values in the interval [0,1]. That is, rather than black & white, we will seek a grey-scale representation.

In the sequel, we formalise the problem by formulating a specific form of ICA model for multivariate binary observations. An early version appears in [15]. We employ a probabilistic framework and make use of the variational methodology to make the inference tractable. Numerical experiments will demonstrate the working of our approach and its advantages over other models of 0–1 data, for the problems considered. Application examples demonstrate the use and the added value of our approach in application areas where ICA methods have not been previously applied/applicable, such as graph or network analysis and DNA fingerprint analysis. A MatLab implementation is available from http://www.cs.bham.ac.uk/~axk/bBICA.m.

1.1. An ndependent factor model with beta sources for binary data

Consider an independent factor model for multivariate i.i.d. binary data $\boldsymbol{x}_n, n = 1,\ldots,N$, where N is the number of observations. A general form of the probability of a datum vector $\boldsymbol{x}_n$ under an independent factor model, in probabilistic terms, is the following:

$$P(\boldsymbol{x}_n) = \int P(\boldsymbol{x}_n|\boldsymbol{b}) \prod_{k=1}^{K} p(b_k)\, \mathrm{d}b_k. \tag{1}$$

Here $b_k, k = 1,\ldots,K$ represent hidden 'source' (component or factor) variables that are assumed to be independent *a priori*, and $\boldsymbol{b} = (b_1,\ldots,b_K)^{\mathrm{T}}$.

The observations are multivariate binary vectors $\boldsymbol{x}_n = (x_{1n},\ldots,x_{tn},\ldots,x_{Tn})^{\mathrm{T}}$ with T samples and N will denote the number of observation (features), $n = 1,\ldots,N$. It is well known from statistics (see e.g. [22]) that the modelling of binary observations requires a distribution that is zero outside the set of the two distinct possible values. Hence, e.g. a Gaussian likelihood model (as employed in most of the previous ICA methods) would not be appropriate in this case and for this reason we employ a conditionally independent Bernoulli likelihood model. This is parameterised by a mean vector that takes the form of a mixture of K components: $\sum_{k=1}^{K} a_{tk} b_{kn}$,

$$P(\boldsymbol{x}_n|\boldsymbol{b}_n) = \prod_{t=1}^{T} \left(\sum_{k=1}^{K} a_{tk} b_{kn}\right)^{x_{tn}} \left(1 - \sum_{k=1}^{K} a_{tk} b_{kn}\right)^{1-x_{tn}}. \tag{2}$$

The conditional independence is a standard assumption in latent variable modelling, meant to force the data dependences to be represented in the latent space. The parameters a_{tk} in (2) are the mixing coefficients of the factor model, and the mixture $\sum_k a_{tk} b_{kn}$ represents the mean parameter of the Bernoulli likelihood.[1] More intuitively, the data $\boldsymbol{x}_n$ is approximated by the combination of factors $\sum_k a_{tk} b_{kn}$, which is indeed the familiar modelling assumption of linear factor models. In both (1) and (2), the conditioning on the parameters a_{tk} is implicit.

The bulk of the design of any factor model, is the specification of the source prior distributions. These determine the statistical characteristics of the sources that we aim to infer. Here we employ independent beta latent prior densities [4]:

$$p(b_k) = B(b_k|\alpha_k^0, \beta_k^0) = \frac{\Gamma(\alpha_k^0 + \beta_k^0)}{\Gamma(\alpha_k^0)\Gamma(\beta_k^0)} (1 - b_k)^{\beta_k^0 - 1} b_k^{\alpha_k^0 - 1}, \tag{3}$$

where α_k^0 and β_k^0 are strictly positive hyperparameters. In the experiments reported, we have set both α_k^0 and β_k^0 to $\frac{1}{2}$, which is the uninformative prior.

The domain of definition of the beta density is $b_k \in [0,1], \forall k$, which is desirable for our purposes, since we may be able to *interpret* the inferred factors as grey-scale representations of the binary data. Interpretability of the components is one of the most important and desirable aspects of independent factor models in general, and this is also what we aim to achieve and exploit in this work. In addition, the particularly flexible shape (see Fig. 1) of the beta density is advantageous for the required density modelling.

The mixing process that we will assume is a convex-linear one, so that the mixing coefficients are all non-negative and satisfy $\sum_k a_{tk} = 1$, for all data-features

[1] For the ease of notations, indices (e.g. in sums or products) are always denoted by small characters and their upper limits by the associated capital letter. Unless indicated otherwise, indices run from 1 to their upper limit.