ICA-based Binary Feature Construction

Ata Kabán[1]* and Ella Bingham[2]

[1] School of Computer Science, The University of Birmingham,
Birmingham, B15 2TT, UK, a.kaban@cs.bham.ac.uk
[2] HIIT BRU, University of Helsinki, Finland, ella@iki.fi

Abstract. We address the problem of interactive feature construction and denoising of binary data. To this end, we develop a variational ICA method, employing a multivariate Bernoulli likelihood and independent Beta source densities. We relate this to other binary data models, demonstrating its advantages in two application domains.

1 Introduction

Binary data becomes more and more abundant, arising from areas as diverse as bioinformatics, e-businesses and paleontological 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.

Previous successes of Independent Component Analysis (ICA) [5] make it an important statistical principle worthy of investigation for tackling such problems. However, contrarily to continuous-valued signals, work on ICA methods for binary data has been very scarce [4, 3]. A few methods exist, though, that seek binary sources [9, 10] from continuous data. Due to the discrete combinatorial nature of the problem, these latter works resort to search heuristics [10] or indeed an exhaustive search [9], that are, at best, computationally intensive.

In this paper we develop a linear ICA model for binary data. We employ a probabilistic framework and make use of the variational methodology to alleviate the computational demand. Application examples will demonstrate the workings of our approach and its advantages over other binary data models.

2 Binary ICA with Beta Sources

Consider an independent factor model for binary data $\boldsymbol{x}$, having a Bernoulli likelihood model and independent Beta latent priors.

$$P(\boldsymbol{x}_n) = \int P(\boldsymbol{x}_n|\boldsymbol{b}) \prod_k p(b_k) db_k \tag{1}$$

$$= \int \prod_t (\sum_k a_{tk} b_k)^{x_{tn}} (1 - \sum_k a_{tk} b_k)^{1-x_{tn}} \prod_k B(b_k|\alpha_k^0, \beta_k^0) db_k \tag{2}$$

[*] Part of this work has been done while visiting HIIT BRU, Helsinki, Finland.

where $B(b|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(1-b)^{\beta-1}b^{\alpha-1}$ is the Beta density [1]. This is defined on the $[0, 1]$ domain, which is desirable for our purposes, since we may be able to interpret the components as grey-scale representations of the binary data. In addition, the particularly flexible shape of the Beta density is advantageous.

Further, a linear-convex mixing process will be assumed, so that the mixing coefficients are all non-negative and satisfy $\sum_k a_{tk} = 1, \forall t = 1 : T$. This is mainly due to computational convenience, since then it follows that $\sum_k a_{tk}b_k$ will necessarily fall into $[0, 1]$ so that we do not need any further nonlinear transformation to obtain the mean parameter of the Bernoulli likelihood. While nonlinear models are also of interest, here we seek the 0-s and 1-s to be exchangeable within the model, and this would not be possible if a nonlinearity is applied to non-negative variables. Thus, a Dirichlet prior may be assumed for the mixing coefficients, to make the specification fully Bayesian.

2.1 Inference and estimation

In order to make the problem tractable, we will employ the well-known Jensen's inequality to lower bound the data probability, and we make use of the factorial posterior approximation to simplify the computations:

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

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

Due to the Bernoulli likelihood term $P(x_{tn}|\boldsymbol{b})$, this integral is still intractable, therefore the ultimate lower bound will be obtained by a further application of Jensen's inequality. The convexity constraint imposed on the mixing proportions comes in useful, as the likelihood term can be rewritten and lower bounded:

$$\log P(x_{tn}|\boldsymbol{b}) = \log \left\{ (\sum_k a_{tk} b_k)^{x_{tn}} (1 - \sum_k a_{tk} b_k)^{1-x_{tn}} \right\}$$
$$= \log \left\{ \sum_k a_{tk} b_k^{x_{tn}} (1 - b_k)^{1-x_{tn}} \right\} \geq \sum_k Q_{k|t,n,x_{tn}} \log \frac{a_{tk} b_k^{x_{tn}} (1 - b_k)^{1-x_{tn}}}{Q_{k|t,n,x_{tn}}} \quad (3)$$

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

Using (3) we obtain a lower bound on the log likelihood, which is tractable and will be referred to as $\mathcal{L}^{bound}$.

2.2 Variational solution

Let $q_n(b_k) = B(b_k|\alpha_{kn}, \beta_{kn})$ be parameterised Beta variational posteriors with variational parameters α_{kn}, β_{kn}. Then, maximising $\mathcal{L}^{bound}$ yields the following update equations for the variational parameters

$$\alpha_{kn} = \alpha_k^0 + \sum_t x_{tn} Q_{k|t,n,x_{tn}=0} = \alpha_k^0 + e^{\langle \log b_{kn} \rangle} \sum_t \frac{x_{tn} a_{tk}}{\sum_k a_{tk} e^{\langle \log b_{kn} \rangle}} \quad (4)$$

$$\beta_{kn} = \beta_k^0 + \sum_t (1 - x_{tn}) Q_{k|t,n,x_{tn}=1} = \beta_k^0 + e^{\langle \log(1-b_{kn}) \rangle} \sum_t \frac{(1 - x_{tn}) a_{tk}}{\sum_k a_{tk} e^{\langle \log(1-b_{kn}) \rangle}} \tag{5}$$

where

$$Q_{k|t,n,x_{tn}} \propto a_{tk} (e^{\langle \log b_{kn} \rangle})^{x_{tn}} (e^{\langle \log(1-b_{kn}) \rangle})^{1-x_{tn}} \tag{6}$$

is obtained by maximising $\mathcal{L}^{bound}$ w.r.t. $Q_{k|t,n,x_{tn}}$ and this has been replaced into the expressions of all variational parameter estimates above.

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

Maximising $\mathcal{L}^{bound}$ in a_{tk} under the constraint that $\sum_k a_{tk} = 1$ and replacing the expression of $Q_{k|t,n,x_{tn}}$ as before, yields the update equation below.

$$a_{tk} \propto a_{tk} \left\{ \sum_n \frac{x_{tn}}{\sum_k a_{tk} e^{\langle \log b_{kn} \rangle}} e^{\langle \log b_{kn} \rangle} + \frac{1 - x_{tn}}{\sum_k a_{tk} e^{\langle \log(1-b_{kn}) \rangle}} e^{\langle \log(1-b_{kn}) \rangle} \right\} \tag{7}$$

Finally, the prior parameters α_k^0 and β_k^0 will both be set to one, in order to express a uniform prior.

To make some connections with earlier work, it can easily be shown that a maximum likelihood estimator for our model (2) would yield equations that (after some algebra) are identical to the aspect Bernoulli (AB) algorithm in [7]. Vice-versa, the above construction offers an interpretation of AB as an ICA model. By analogy, other popular aspect models [2, 3] may also be related to ICA in a similar manner, and this is different from, and complementary to the connection initially envisaged in [3].

2.3 Bayesian model selection

As already mentioned, a prior may also be naturally specified for the mixing coefficients, and due to the imposed convexity constraint, a Dirichlet is appropriate. As a result, the optimal number of components can determined simply by choosing the model order that maximises the log of the data evidence bound

$$E_{q_t(\boldsymbol{a})}[\mathcal{L}^{bound}] + E_{q_t(\boldsymbol{a})}[\log Dir(\boldsymbol{a}|\boldsymbol{\gamma}^0)] - E_{q_t(\boldsymbol{a})}[\log q_t(\boldsymbol{a})] \tag{8}$$

where $q_t(\boldsymbol{a}) = Dir(\boldsymbol{a}|\boldsymbol{\gamma}_t)$ is the variational posterior of the mixing variable.

The modification this brings to the previously presented estimation procedure is minimal — denoting by γ_{tk} the additional variational parameters associated with a_{tk} and omitting the straightforward algebra, the parameters a_{tk} in (4) will need to be replaced by $e^{\langle \log \gamma_{tk} \rangle}$ and instead of eq (5) we will have:

$$\gamma_{tk} = \gamma_k^0 + e^{\langle \log a_{tk} \rangle} \left\{ \sum_n \frac{x_{tn} e^{\langle \log b_{kn} \rangle}}{\sum_k e^{\langle \log a_{tk} \rangle} e^{\langle \log b_{kn} \rangle}} + \frac{(1 - x_{tn}) e^{\langle \log(1-b_{kn}) \rangle}}{\sum_k e^{\langle \log a_{tk} \rangle} e^{\langle \log(1-b_{kn}) \rangle}} \right\} \tag{9}$$

The parameter of the Dirichlet prior, γ_{tk}^0 will again be set to 1 to express a uniform prior, and the remaining posterior expectation in (9) is computed as $\langle \log a_{tk} \rangle \equiv E_{q_t(\boldsymbol{a})}[\log a_k] = \psi(\gamma_{tk}) - \psi(\sum_{k'} \gamma_{tk'})$.

3 Analyst input and posterior data reconstruction

Perhaps the greatest reason for the popularity of ICA methods for exploratory data analysis is that the independent components are often easier to comprehend and interpret by humans separately, rather than in their mixture. This has been exploited in numerous applications, most notably for signal denoising [6]. Once the independent signals of different genuine and artifact sources are separated from the data, artifact-corrected signals may be derived by eliminating the contributions of the artifact sources. Our methodology is conceptually similar, although the formalism differs according to our probabilistic framework.

Let us denote the posterior means obtained from our algorithm by $\langle a_{tk} \rangle$ and $\langle b_{kn} \rangle$: $\langle b_{kn} \rangle = E_{q_n(b_k)}[b_k] = \int db_k b_k B(b_k|\alpha_{kn}, \beta_{kn}) = \frac{\alpha_{kn}}{\alpha_{kn}+\beta_{kn}}$ and analogously $\langle a_{tk} \rangle = \frac{\gamma_{tk}}{\sum_{k'} \gamma_{tk'}}$. These are themselves discrete probabilities, so that $\sum_k \langle a_{tk} \rangle = 1$. After inspecting the independent components $\langle \boldsymbol{b}_k \rangle$, the elimination of undesired components may now be accomplished by specifying a probability value, $P(u|k)$, for each component and using these to modify our unsupervised estimates. Denoting by $P_t(k)$ the posterior expectations $\langle a_{tk} \rangle$, for each t, the Bayes rule will provide us the post-processed data representation.

$$\langle a_{tk} \rangle_{postproc} := P_t(k|u) = \frac{P_t(k)P(u|k)}{\sum_{k'} P_t(k')P(u|k')} \tag{10}$$

Typically a 0 probability will be specified for components that are capturing undesirable noise, while 1 will specify a clearly meaningful component. Clearly, if for a component k a value of $p(u|k) = 0$ was specified, then $\langle a_{tk} \rangle_{postproc} = 0$ will become zero for all t. Naturally, the formalism straightforwardly permits the specification of analyst inputs at more detailed levels. E.g. nothing prevents us from specifying a separate set of probabilities, $P(u|k,t)$, for each t. However, we may typically expect human experts to feed back on the components' level, since those are hoped to provide some interpretable representations.

For computing the posterior data reconstruction, we re-express the above in terms of a conditional posterior: $q_t(\boldsymbol{a}|\boldsymbol{u}) := Dir(\boldsymbol{a}|\boldsymbol{\gamma}_t \circ P(u|.))$, whose expectation is exactly $\langle a_{tk}|\boldsymbol{u} \rangle = \langle a_{tk} \rangle_{postproc}$. Here, $\circ$ denotes element-wise product and $\boldsymbol{u}$ is the random vector of $u|k$ when $k = 1 : K$. Then the posterior post-processed data reconstruction can be computed as follows (omitting the algebra):

$$P(\hat{x}_{tn}|\boldsymbol{X}, \boldsymbol{u}) = \int d\boldsymbol{a} d\boldsymbol{b} P(\hat{x}_{tn}|\boldsymbol{a}, \boldsymbol{b}) q_t(\boldsymbol{a}|\boldsymbol{u}) \prod_k q_n(b_k) \tag{11}$$

$$= (\sum_k \langle a_{tk}|\boldsymbol{u} \rangle \langle b_{kn} \rangle)^{\hat{x}_{tn}} (1 - \sum_k \langle a_{tk}|\boldsymbol{u} \rangle \langle b_{kn} \rangle)^{1-\hat{x}_{tn}} \tag{12}$$

In consequence, the grey-scale posterior reconstruction of the (t, n)-th data entry is

$$\langle \hat{x}_{tn}|\boldsymbol{u} \rangle = p(\hat{x}_{tn} = 1|\boldsymbol{X}, \boldsymbol{u}) = \sum_k \langle a_{tk}|\boldsymbol{u} \rangle \langle b_{kn} \rangle \tag{13}$$

and so the binary reconstruction is given by thresholding this value.

4 Experiments

4.1 Restoration of corrupted binary images

For the first set of experiments we use a data set of handwritten digit images[3]. The subset of the first five digits were taken, each having 200 examples, which totals 1,000 image instances. We artificially created a corrupted version of this data set, by simulating a uniformly varying process of degradation, which turns off some of the pixels that were initially 'on'. Fifteen randomly chosen examples are shown from the initial data set, along with their corrupted version, on Figure 1. Figure 2 then shows the ICA representation obtained: several components can clearly be recognised as typical digits, and one other – completely blank – separates out the corruption factor. Inspecting the mixing proportions for the data instances shown earlier, it is clear that the white component is indeed present in those images that suffered a degradation. To remove the noise component, we apply the procedure described earlier. The results can be followed on Figure 3: The grey-scale posterior reconstruction of the data has indeed filtered out the degradation source and presents a smoothed reconstruction of the initial clean data. The grey levels correspond to probabilities of pixels being 'on'. Thresholding these probabilities at 0.5 gives us the binary reconstruction of the data shown on the right-hand plot. The degradation has now been eliminated.

A comparative set of experiments has then been conducted in order to objectively and quantitatively assess the performance of our method in reconstructing the clean data from its corrupted version. We included a comprehensive set of binary data analysis methods in this comparison: mixtures of Bernoulli (MB), Bernoulli (logistic) PCA [11] (LPCA), our binary ICA with and without post-processing (BICA-postproc and BICA respectively), and a Bernoulli version of non-negative matrix factorisation [8], that we created for the purpose of this comparison (BNMF). (For the latter, a shifted and rescaled sigmoid nonlinearity was used, which transforms the non-negatively constrained factors and mixing proportions into the [0,1] interval.) None of the methods except BICA was able to separate out the noise factor. In consequence no obvious correction post-processing is applicable to the other methods. In this experiment, 500 corrupted images were used for training and another 500 corrupted images formed an independent test set. For the previously unseen data instances, the required posteriors were first estimated. In the case of BNMF we just implemented a Maximum Likelihood estimation method and in this case the required parameter matrix was estimated anew for the previously unseen test data. The upper plots of Figure 4 show the areas under the ROC curve of the posterior data reconstruction (both grey-scale and black&white, using a threshold of 0.5), averaged over all pixels of the corrupted test set. LPCA is the overall winner in reconstructing the corrupted test data set. The lower plots of the same figure, in turn, show the AUC values averaged over the blank pixels of the test images, but computed against the pixel values of the true, uncorrupted test set (not used anywhere else). As we can see, the proposed post-processing, by the removal of

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

the automatically separated noise component, BICA becomes the most successful in this exercise – comparable with the nonlinear and time-consuming LPCA at grey-scale reconstruction and net superior at binary reconstruction.

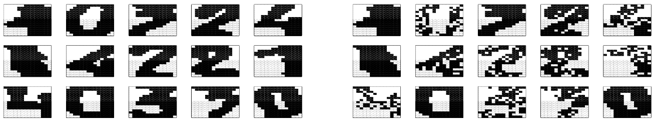

Fig. 1. Examples of clean (left) and corrupted (right) images.

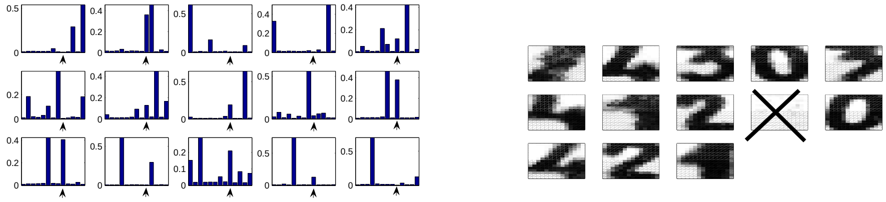

Fig. 2. Right: Independent components estimated from the corrupted binary image data set; Left: The mixing coefficients associated with the examples shown on the right hand plot of Fig.1. Small arrow heads point to the mixing coefficients associated with the noise component.

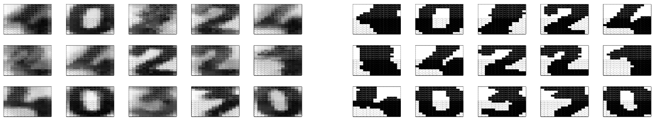

Fig. 3. Reconstructed grey-scale (left) and binary (right) images after the post-processing.

4.2 Age discovery and missingness detection in paleontological data

We now demonstrate our method in paleontological data[4]. The data consists of findings of 139 mammals among 501 sites of excavation and is seen in Figure 5 (leftmost plot). Four components have been estimated, out of which three turned out to capture contiguous disjoint time periods. The fourth component in turn is completely blank — having all elements nearly zero. The second left plot of Figure 5 shows the box plots of the ages of remains[5], weighted by b_{kn}. The

[4] NOW database, http://www.helsinki.fi/science/now/, a public resource based on collaboration between mammal paleontologists

[5] The age information is auxiliary and it is not used during the parameter estimation.

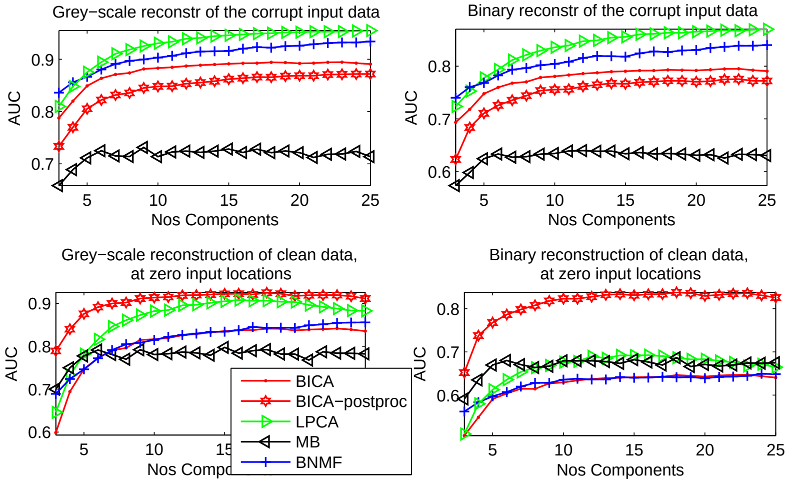

Fig. 4. Comparison of BICA with other binary data models on test inputs. Since both the training and the test data sets are corrupted, all methods try to reconstruct the data including the corruptions, LPCA being the best (upper plots). However, by the described post-processing, BICA stops reconstructing the corrupted regions, instead it becomes net superior in terms of restoration of the uncorrupted images (lower plots).

Kolmogorov-Smirnov test indicates that these distributions are indeed distinct: the P values range between $5 \cdot 10^{-13}$ and $3 \cdot 10^{-4}$. The blank component is the one shrunken to zero on this figure – clearly it does not contribute to the age discovery. In turn, its presence indicates that not all zero observations are due to age, but another reason for absence of remains exists.

Often, remains of a mammal are not observed at a site even though it probably lived there, as the preservation, recovery and identification of fossils are subject to random effects. According to palaeontologists[6], an indication of missingness can be derived from the age order of the sites: if a mammal is observed at two sites but not at an intermediate site, it is possible (although not certain) that an observation at the intermediate site is missing. This may be the additional independent noise factor that our method has separated out, and in order to verify this, we will now remove this noise factor from the data. Employing the probabilistic post-processing procedure described previously, and thresholding at 0.5 (see Figure 5, third plot from the left), we obtain a significant decrease in such intermediate, ”probably missing” values: 1369 of them will be filled in. Furthermore, by thresholding at a smaller value of 0.3481 (obtained by considering all such intermediate values as missing, and dividing the number of 1s plus missing values by the size of the data) the decrease in ”probably missing” values raises to 3642. The continuity of mammals as recovered by our binary ICA is now quite apparent on the rightmost plot of Figure 5.

[6] Professor Mikael Fortelius, University of Helsinki, personal communication.

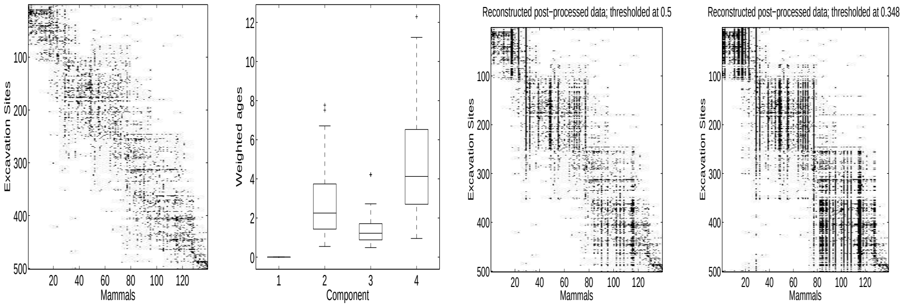

Fig. 5. From left to right: The palaeontological data, both the sites and the remains of mammals are ordered by age, for the ease of visual analysis of the results; Distributions of ages of mammals, weighted by $\langle b_{kn} \rangle$, for each component; Binary reconstruction of the absences in the data after having removed the noise component, using a threshold of 0.5 – these are superimposed with the observed presences; Binary reconstruction, when using an estimated threshold.

5 Conclusions

We have devised a variational ICA method for binary data, employing independent Beta latent densities. This turned out to be a flexible model and has allowed us to include human input in a principled manner. We demonstrated the use of our approach on two application examples.

References

1. J.M Bernardo and A.F.M Smith. Bayesian Theory. Wiley, 2001.
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 and A Jakulin. Applying Discrete PCA in Data Analysis. Proc. 20th Conference on Uncertainty in Artificial Intelligence, pp. 59 - 66, 2004.
4. J Himberg and A Hyvärinen. Independent Component Analysis for Binary Data: an Experimental Study. Proc. ICA2001, pp. 552-556, 2001.
5. A Hyvärinen, J Karhunen, E Oja. Independent Component Analysis. Wiley, 2001.
6. T.P Jung, S Makeig, C Humphries, T.W Lee, M.J McKeown, V Iragui, T.J Sejnowski. Removing Electroencephalographic Artifacts by Blind Source Separation, Psychophysiology, 37:163-78, 2000.
7. A Kabán, E Bingham, T Hirsimäki, Learning to Read Between the Lines: The Aspect Bernoulli Model, Proc. SIAM Int Conf on Data Mining, 2004, pp.462–466.
8. D.D Lee and H.S Seung. Algorithms for non-negative matrix factorization, Advances in Neural Information Processing Systems, 2000, pp.556–562.
9. P. Pajunen. Blind Separation of Binary Sources With Less Sensors Than Sources. Proc Int. Conf on Neural Networks (ICNN-97), 1997, pp. 1994-1997.
10. E Segal, A Battle, D Koller. Decomposing gene expression into cellular components. Proc. Pacific Symposium on Biocomputing, 2003, pp. 89–100.
11. M.E Tipping. Probabilistic visualisation of high dimensional data, Advances in Neural Information Processing Systems, 1999, pp. 592–598.