BATM: Bernoulli Aspect Topological Model

Rodolphe Priam Mohamed Nadif

Abstract

The mixture models behave very well to cluster large samples of continuous or categorical data. Adding a vicinity constraint permits them to project data like factorial methods but in a nonlinear way. In this paper we present a new model called *Bernoulli Aspect Topological Mapping* (BATM) : a generative self-organizing map to deal with binary data by a new automatic map smoothing and an original initialization.

1 Introduction

The visualization of the main correlations and similarities of a data sample is the goal pursued by the factorial data analysis methods [1]. These methods often seek orthogonal informative directions in the data cloud with most of the projected variance as inertia is carriering sense. A well driven decomposition of the inertia into projective planes explains which data points are near each other and why they are, i.e. which variables are concerned and how are strong their local correlations. Although these methods are very powerful, large data samples need new efficient methods. In this context, the Kohonen maps [2] are well known in the visual data analysis field, they generalize the factorial methods such as the Principal Component Method (PCA) [1] for continuous data. More generally, Self-Organizing Maps (SOM) [2] are clustering methods with a vicinity constraint on the classes to give a topological sense to the obtained partition. So, class centers are drawn on the plane with the data which are the nearer. The Generative Topographic Mapping (GTM) [3] is a probabilistic Self-Organizing Map with constrained means for continuous data, but it is not relevant for categorical or binary data. To this end, some recent models [4] [5] [6] were introduced by generalizing GTM to the self-organization of the classical asymetric discrete mixture models. Moreover, Hofman and Puzicha have proposed the symetric aspect model approach [7] which treats to classify simultaneously the rows and the colums of a data matrix. This approach is beneficial in different domains such as text mining, image segmentation [7]. Hofman has proposed a topological aspect model [8] for contingency table. In this paper, we study a new version for binary data by proposing a new map smoothing and an original initialization to accelerate the convergence of our method. The probabilities are properly parametrized like the GTM to induce a self-organization of the latent factors which bring to us a new way to visualize discrete datasets of zero or one multidimensional vectors.

The paper is organized as follows. Section 2 begins with the development of the model and its estimation by maximizing the loglikelihood. In Section 3, we focus on the experiments to validate our model. An application on a binary dataset often used as a benchmark and an illustrative experiment for textual data are presented. Finally, Section 4 summarizes the main points of this paper and future works in progress.

2 BATM model

The proposed model supposes as [9] independance of the $I \times J$ cells $x_{ij} \in \{0,1\}$ from a binary matrix, by modelling each unidimensional probability of the x_{ij} observed as a mixture of K Bernoullian laws: $Pr(x_{ij} = 1) = \mathbb{E}[x_{ij}] \propto \sum_k \pi_{ki} a_{jk}$ with π_{ki} the mixing probabilities such as $\sum_k \pi_{ki} = 1$. For the dataset $\mathcal{D} = \{x_i\}_{i=1}^{i=I}$ where $x_i = (x_{i1}, x_{i2}, \cdots, x_{iJ})$ is a datum, the corresponding loglikelihood is:

$$\mathcal{L}(\theta|\mathcal{D}) = \sum_i \sum_j \log \Big[\sum_k \pi_{ki} a_{jk}^{x_{ij}} (1 - a_{jk})^{1-x_{ij}} \Big]$$

To induce a topological ordering to the probabilities, we consider the K coordinates $\{s_k\}_{k=1}^{k=K}$ from a regular bidimensional mesh which models a discretized plane where the dataset is going to be layed out. The mesh is projected in a higher space of L dimensions by a nonlinear transformation $\xi_k = \Phi(s_k) = (\phi_1(s_k), \phi_2(s_k), \cdots, \phi_L(s_k))$ with L basis functions ; we write $\Phi = (\xi_1^T|\xi_2^T|\cdots|\xi_K^T)^T$. The a_{jk} are then seen as the knots of a discrete nonlinear surface: the Bernoulli laws are parametrized as a sigmoid function $a_{jk} = \sigma(w_j^T \xi_k)$ where $\sigma(u) = e^u/(1+e^u)$ is the logistic function. The loglikelihood becomes:

$$\mathcal{L}(\theta|\mathcal{D}) = \sum_{i,j} \log \Big[\sum_k \pi_{ki} \sigma(w_j^T \xi_k)^{x_{ij}} (1 - \sigma(w_j^T \xi_k))^{1-x_{ij}} \Big]$$

This model is a binary version of the Probabilitic LSA or pLSA [7] where its Multinomial hypothesis is replaced with a constrained Bernoulli law on each cell of the data matrix $A = [x_1^T|x_2^T|\cdots|x_I^T]^T$. Unknow parameters are estimated in the following section.

2.1 GEM estimation Inference of this model is done by maximizing the loglikelihood which is intractable in an exact

closed form solution because of the non linearities from the sigmoid functions. So we study the gradient-based approach of the Generalized Expectation-Maximization (GEM) algorithm from Dempster et al. [10] wich permits to deal with the log of a sum. This approach assumes the likelihood completed by the knowledge of the partition $\mathcal{Z} = \{\mathcal{Z}_1, \mathcal{Z}_2, \cdots, \mathcal{Z}_K\}$:

$$\mathcal{L}(\theta, Z|\mathcal{D}) = \sum_i \sum_j \log\left[\pi_{z_i i}\, a_{jz_i}^{x_{ij}} (1 - a_{jz_i})^{1-x_{ij}}\right]$$

with z_i the latent variable whose unknown value is in $\{1, 2 \cdots, K\}$. The EM [11] algorithm is based on the maximization of the conditionnal mean of the complete loglikelihood given the data and the parameters of the preceeding iteration. Having $P^{(t)}(Z|\mathcal{D})$ from the t-st step, we maximize in step $t+1$:

$$\begin{aligned}\mathcal{Q}(\theta|\theta^{(t)}) &= \mathbb{E}_{P^{(t)}(Z|\mathcal{D})}[\mathcal{L}(\theta, Z|\mathcal{D})] \\ &= \sum_{i,j,k} P^{(t)}_{k|i,j,x_{ij}} \Big\{ \log \pi_{ki} \\ &+ x_{ij}\xi_k^T w_j - log(1 + exp(\xi_k^T w_j))\Big\}\end{aligned}$$

with: $P_{k|i,j,x_{ij}} \propto \pi_{ki} a_{jk}^{x_{ij}} (1 - a_{jk})^{1-x_{ij}}$ the posterior probability that x_{ij} was generated by the k-st aspect. A closed form for maximizing this quantity doesn't exist yet, so we use a gradient approach to calculate $w^{(t+1)} = argmax_w Q(\theta|\theta^{(t)})$, while direct derivation gives us:

$$\pi_{ki}^{(t+1)} = argmax_{\pi_{ki}} Q(\theta|\theta^{(t)}) = \sum_j P^{(t)}_{k|i,j,x_{ij}}/J$$

By derivating the criterion, we get the Gradient vector $\mathbf{Q}_j^{(t)}$ and the Hessian matrix $\mathbf{H}_j^{(t)}$. As the Hessian is a block diagonal matrix, we are able to increase the loglikelihood with a Newton-Raphson ascent step:

$$w_j^{(t+1)} = w_j^{(t)} - {\mathbf{H}_j^{(t)}}^{-1} \mathbf{Q}_j^{(t)}$$

Iterating $\pi_{ki}^{(t+1)}$ and $w_j^{(t+1)}$ converge to a maximum of $\mathcal{L}(\theta|\mathcal{D})$ that we write $\hat{\theta}$. To avoid overfitting and bad numerical solutions, we add a bayesian [12] gaussian prior: $\mathcal{Q}(\theta|\theta^{(t)}) - \alpha \sum_j w_j^T w_j$. The correction of the estimates is done by adding: $-\alpha w_j$ to the gradient $\mathbf{Q}_j$ and $-\alpha \mathbb{I}_L$ to the diagonal of the Hessian $\mathbf{H}_j$. The value of the hyperparameter α is most of the time manually chosen in the litterature as 0.01 for instance.

2.2 IRLS formulation We write the Newton-Raphson process in a matrix form which sounds like an *Iteratively Reweighted Least Squares* (IRLS [13]) step. For j from 1 to J:

$$\begin{aligned}\mathbf{Q}_j^{(t)} &= \Phi^T\left[R_j^{(t)} A_j - G_j^{(t)} a_j^{(t)}\right] - 0.01 w_j^{(t)} \\ \mathbf{H}_j^{(t)} &= -\Phi^T G_j^{(t)} F_j^{(t)} \Phi - 0.01 \mathbb{I}_L\end{aligned}$$

We have $R_j^{(t)}$ the $K \times I$ matrix with a posteriory probabilities $P^{(t)}_{k|i,j,x_{ij}}$ as cell values, the matrix $G_j^{(t)}$ is the diagonal matrix with $\sum_i P^{(t)}_{k|i,j,x_{ij}}$ as non null elements, A_j is the j-st column of A, $a_j^{(t)}$ is a column vector with the $a_{jk}^{(t)}$ as j-st component, $F_j^{(t)}$ is the diagonal matrix with $a_{jk}^{(t)}(1 - a_{jk}^{(t)})$ on its diagonal, and $\mathbb{I}_L$ is the $L \times L$ identity matrix.

To numerically accelerate the algorithm, a Bohning [14] approach replaces the exact Hessian quite heavy to calculate by one fixed matrix. For instance $\mathbf{B} = -\frac{I}{4}\Phi^T\Phi - 0.01\mathbb{I}_L$ which is such as $\mathbf{H}_j^{(t)} \succeq \mathbf{B}$, i.e. $\mathbf{H}_j^{(t)} - \mathbf{B}$ is a non-negative definite, symetric matrix, and so we are still maximizing the likelihood. This matrix gave a slow convergence so we propose a variational alternative algorithm with a closed form maximization step contrary to the preceeding generalized EM scheme.

2.3 Variational estimation Following the bound[1] [15] for $\log(1 + exp(\xi_k^T w_j))$ we get the new criterion:

$$\begin{aligned}\tilde{\mathcal{Q}}(\theta|\theta^{(t)}) &= \sum_{i,j,k} P^{(t)}_{k|i,j,x_{ij}} \Big\{ \log \pi_{ki} + (x_{ij} - 0.5)\xi_k^T w_j \\ &+ \lambda(\epsilon_j)[(\xi_k^T w_j)^2 - \epsilon_j^2] + 0.5\epsilon_j - \log(1 + exp(\epsilon_j))\Big\}\end{aligned}$$

with $\lambda(\epsilon_j) = -tanh(0.5\epsilon_j)/(4\epsilon_j)$ such as $\mathcal{Q}(\theta|\theta^{(t)}) \geq \tilde{\mathcal{Q}}(\theta|\theta^{(t)})$ where ϵ_j is a variational parameter to be found by maximizing $\tilde{\mathcal{Q}}$. By derivating this new criterion we get the new Maximization step :

$$\begin{aligned}\epsilon_j^{(t)} &= \sqrt{\frac{{w_j^{(t)}}^T \Phi^T G_j^{(t)} \Phi w_j^{(t)}}{I}} \\ w_j^{(t+1)} &= \Big[-2\lambda(\epsilon_j^{(t)})\Phi^T G_j^{(t)} \Phi - 0.01\mathbb{I}_L\Big]^{-1} \Phi^T R_j^{(t)} A'_j\end{aligned}$$

where A'_j is the column vector with $x_{ij} - 0.5$ as i-st components. Finally, we provided three main algorithms -and a fourth one in the next section- to estimate the parameters of the model. While discarding the ineffective simple gradient steps, the IRLS algorithm gives the best loglikelihood in our case as the experiments show in the next section.

[1] $\log\sigma(u) \geq u/2 + \lambda(\epsilon)(u^2 - \epsilon^2) + \log\sigma(\epsilon) - \epsilon/2$ for the concavity reasons.

3 Simulations

3.1 Initialization of the model Random trials are a solution to local minima where an optimization algorithm is trapped. One way to get the best convergence is to have a good initialization. As Kohonen maps are generalizing Principal Component Analysis, the first plane from this method provides [16] an appealing first point for the parameters. Let's have (X_i^c, Y_i^c) the continuous coordinates on the first plane from a factorial projection as PCA [17], CA [18], LSA [19] or even those from a nonlinear mapping like MDS [20]. Then, a regular mesh is drawn on this first projection with each cell symbolizing a factor from the BATM model, and so, x_i is put in the $z_i^{(0)}$-st cell where it falls into for this initial plane. We initialize mixing probabilities as $\pi_{ik}^{(0)} \propto h(k, z_i^{(0)})$ for a smoothing function like the vicinity one from the Kohonen's map, i.e. $h(k, z_i^{(0)}) \propto exp(-||s_k - s_{z_i^{(0)}}||^2/\sigma)$ for σ well chosen. Then:

$$a_{jk}^{(0)} = \frac{\sum_i \pi_{ik}^{(0)} x_{ij} + \alpha}{\sum_i \pi_{ik}^{(0)} + I\alpha}$$

for $\alpha > 0$ well chosen whose goal is to avoid empty cells. Finally, $LP^{(0)}$ is the $K \times J$ matrix with cell values equal to $\log[a_{jk}^{(0)}/(1-a_{jk}^{(0)})]$. This matrix permits us to find the initial $W^{(0)}$ matrix which column-aggregates the $w_j^{(0)}$ vectors, as a the regression solution on the matrix Φ:

$$W^{(0)} = [w_1^{(0)}|w_2^{(0)}|\cdots|w_J^{(0)}] = (\Phi^T\Phi)^{-1}\Phi^T LP^{(0)}$$

Contrary to a self-organizing map for continuous data, we have to deal with binay data where the continous coordinates from the first factorial plane cannot be used as initial values for our center classes. So, we have constructed those center by clustering this coordinates and smoothing their hard affectations.

3.2 Topological organization of the rows It can be interesting to add a constraint on the rows to help convergence towards a well organized state and accelerate the algorithm. As a soft-max [21] solution appears heavy we propose a lighter solution by adding a penalty term from the TNEM [22] approach. Roughly speaking, the idea behind this algorithm is to cluster the data vectors with a spatial smoothing of the aspects components from the BATM model, as a Hidden Random Field Model [23] [24] [25] does. This is written here:

$$\tilde{\mathcal{Q}}(\theta|\theta^{(t)}) \quad = \quad \mathcal{Q}(\theta|\theta^{(t)}) + \frac{\beta}{2}\sum_i \pi_i^T \mathbf{V}\pi_i$$

where π_i is the vector of the π_{ki} as components, and $\mathbf{V}$ is the neighboorhood matrix from the self-organizing map, i.e. $V_{k\ell} = h(k, \ell)$, eventually replaced by the binary adjacency matrix of the map, i.e. $V_{k\ell} = 1$ iff k is near ℓ. This new step is written:

$$\pi_{ki}^{(t+1)} = \frac{\sum_j P_{k|i,j,x_{ij}}^{(t)} + \beta\,\pi_{ki}^{(t+1)} \sum_\ell V_{k\ell}\,\pi_{\ell i}^{(t+1)}}{J + \beta\,{\pi_i^{(t+1)}}^T \mathbf{V}\pi_i^{(t+1)}}$$

which is solved by iterating this equality and reinjecting in the right member old current values until convergence. We obviously retrieve the nonconstrained estimation when β is zero. As we discard the entropy term, we end to a new constrained algorithm called TNEM2, more general than the original TNEM one: it can be applied to any model with probabilities to induce thei topographic smoothing. An alternative topological constrained clustering of the rows is to add a parametrization from the GTM approach for the probabilities π_{ki}. The function to be maximised on the rows is then:

$$\mathcal{Q}_I(\theta|\theta^{(t)}) = \sum_{i,j,k} P_{k|i,j,x_{ij}}^{(t)} \left\{ \xi_k^T w_i - \log \sum_\ell exp(\xi_\ell^T w_i) \right\}$$

where the w_i are the new unknown parameters that we are seeking for. We do a Newton-Raphson ascent as in the preceeding part, by looping over the I rows of the data matrix, and calculating the gradient vector $\mathbf{Q}_i^{(t)} = J\Phi^T(\pi_i^{(t+1)} - \pi_i^{(t)})$ and the Hessian matrix $\mathbf{H}_i^{(t)} = -J\Phi^T(F_i - \pi_i^{(t)}{\pi_i^{(t)}}^T)\Phi$ where $F_i^{(t)}$ is the diagonal matrix with $\pi_{ki}^{(t)}$ as non zero cell values. The parameters are initialized to $w_i^{(0)}$, as before, by using a regression step over the new matrix $LP_{\mathcal{I}}^{(0)}$ whose cells are the logarithm of the probabilities $\pi_{ik}^{(0)}$. Finally, we end to a fourth algorithm IRLS+TNEM2 with the IRLS loop for the columns and the TNEM2 step for the rows. We explain in the next section how this process behaves well in practice.

3.3 Post-processing of the final map The final map shows a mesh of spacially well organized class centers where one can place each data sample in its nearer center. For classical self-organizing maps one uses an Euclidian distance between the center vector and the data vector. Here, the model permits a probabilistic alternative as we have the probability that a data was generated by an aspect, so each data x_i can be affected to the maximum a posteriori (MAP) center, i.e. $\hat{z}_i = argmax_k \hat{\pi}_{ki}$. In the same way, each variable, corresponding to the j-st component, can be affected to the center of label $\hat{z}_j = argmax_j \hat{a}_{jk}$. So it induces the bidimensional positions $p_i = s_{\hat{z}_i}$ and $p_j = s_{\hat{z}_j}$. An other way to use the final map is to project each data sample as a mean value [3] instead of the preceeding MAP value. This means the positions $\tilde{p}_i = \sum_k \hat{\pi}_{ki} s_k$ and $\tilde{p}_j = \sum_k (\hat{a}_{jk}/\sum_\ell \hat{a}_{\ell j}) s_k$.

3.4 Experiments We experiment our model on several database to validate our approach. For exemple, on the Zoo

Figure 1: A BATM 8×8 table for the Zoo dataset.

database[2] which counts 101 animals with seven classes of species and 21 binary re-coded characteristics. The BATM method ends to a well organized map on Figure 1 where we recognize the seven classes that BATM has projected on its map. The table colouring is done by hand with the help [26] from a hierarchical clustering algorithm [27]. The loglikelihood from the four presented algorithms is given on Figure 3 for the Zoo dataset, demonstrating the superiority of the classical IRLS compared to more recent algorithms in alternative. However overfitting can lead to a not enought smooth solution and we prefer using the IRLS+TNEM2 approach because of its efficiency despite its smaller loglikelihood. Its lower value comes from the penalization term which helps a quick auto-organization of the rows as can been seen here when it stops very early before the three other given algorithms for a same stopping criterion (relative loglikelihood below the threshold 10E-5). The initialization step from a Correspondence Analysis [18] is illustrated on Figure 2, demonstrating the interest of a self-organizing map: while the linear factorial method is not able to show the seven classes on the first factorial plane, our BATM map extracts theses classes and finds there statistical links thanks to the vicinity property. A textual dataset is projected by the BATM method. The dataset is a sample from the Classic3 [28] file which counts three classes from MEDLINE, CISI, CRANFIELD. We select 450 documents

[2] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/zoo/zoo.names

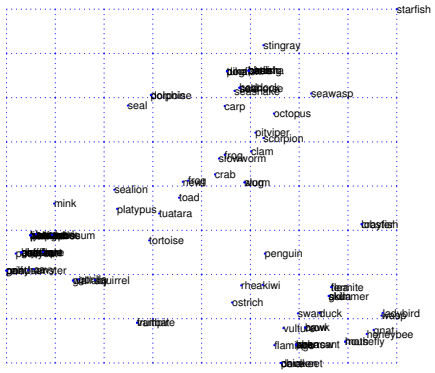

Figure 2: Initialization of the BATM map for the Zoo dataset.

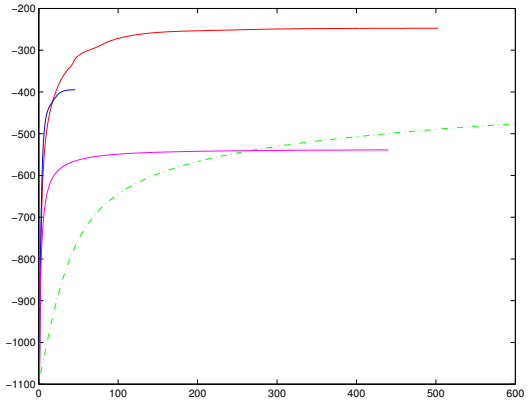

Figure 3: The four loglikelihood curves from the IRLS, IRLS+TNEM2, Bohning (end not shown) and variational gradient are drawn from the top to the bottom of the graph.

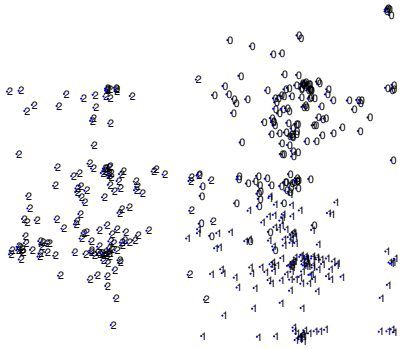

Figure 4: Mean projection for the Classic3 sample.

from this file, by randomly drawing 150 documents from each class. Then we select the more frequent words over 30 from all the vocabulary of 4303 terms and we end to a matrix with approximately 450 rows and 170 columns while discarding the empty rows. We show the mean positions for the label of the corresponding documents in Figure 4. We are able to see the three classes almost well separated by our non linear mapping.

4 Conclusion and discussion

We have presented a new self-organizing map model for binary data as can be found in the image or textual domain. New results for the initialization of a generative mapping method of qualitative data was introduced too. We are now working on the model to deal with bigger matrix. Adding a topological constraint or an entropy scheme on the mixing probabilities should avoid the initialization step. Finally, an alternative to the BATM model is for instance choosing[3] a new $\mathbb{E}(x_{ij})$. The BATM model shall be extended to other data type too, as it is proposed in the appendix section. Estimation can also be improved by finding the best hyperparameter β or working on the variational formulation. The first interest is to obtain an alternative to a more classical mixture model by a latent variable formulation which leads us a new point of view to understand the contents of the data. To conclude, the simultaneous clustering of the rows and the columns from a numerical data matrix is efficiency performed by a recent generative *Block Mixture model* [29, 30], so we are currently extending this model to the projection of discrete data. One main perspective of our approach is to draw well defined and well readable non linear biplots for large datasets.

Appendix

When the data matrix is a contingency matrix, the Bernoulli law is no longer valid, and a Multinomial or Poisson hypothesis is generally taken. The soft-max parameter is then introduced to deal with multinomial law when the probabilities are constrained as in a topological ordering. This is written in our case $p_{j|k} = e^{w_j^T \xi_k} / \sum_{j'} e^{w_{j'}^T \xi_k}$ with $\sum p_{j|k} = 1$. So, it induces the inversion of a full Hessian when optimization is processed. No variational approach exists to resolve the bottleneck. So, we propose a new way to deal with multinomial, by providing a simple trick. The main idea is to retrieve our nonconstrained parameters by writting $p_{j|k}$ as a joint Bernoulli law with new unknow parameters. This is nothing else that supposing that this probability is the one of the corresponding j-st column from the data matrix where positive values are now one, and each component is independently drawn as a Bernoulli random variable. The following expression $p_{j|k} = p_{jk} \prod_{j' \neq j} (1 - p_{j'k}) \simeq p_{jk}$ with $p_{jk} \in [0, 1]$ gives a valid solution to the multinomial estimation when probabilities are all enought small ; as we have :

$$\begin{aligned} p_{jk}^{(t+1)} &= argmax_{p_{jk}} \sum_i \sum_j \sum_k p_{kij}^{(t)} x_{ij} \log[p_{j|k}] \\ &= argmax_{p_{jk}} \sum_i \sum_j \sum_k p_{kij}^{(t)} x_{ij} \log \left[p_{jk} \prod_{j' \neq j} (1 - p_{j'k}) \right] \end{aligned}$$

we are enable to retrieve the classical expression for multinomial parameters, $p_{jk}^{(t+1)} = \frac{\sum_i p_{kij}^{(t)} x_{ij}}{\sum_i \sum_{j'} p_{kij'}^{(t)} x_{ij'}}$ for a posteriori probabilities $p_{kij}^{(t)}$ and inducing their automatic normalization. As no more constraint is nedeed on every p_{jk} the sigmoid parametrisation $p_{jk} = \sigma(\xi_k^T w_j)$ is available for a topological ordering of our new parameters, without soft-max parameters, providing a new IRLS formula for multinomial laws.

[3] We have for instance: $\mathbb{E}(x_{ij}) = \sum_k \pi_k \pi_{i|k} b_{jk}^{x_{ij}} (1 - b_{jk})^{(1-x_{ij})}$ or $\mathbb{E}(x_{ij}) = \sum_k \pi_k a_{ik}^{x_{ij}} (1 - a_{ik})^{(1-x_{ij})} b_{jk}^{x_{ij}} (1 - b_{jk})^{(1-x_{ij})}$.

References

[1] L. Lebart, A. Morineau, and K. Warwick, *Multivariate Descriptive Statistical Analysis*. J. Wiley, 1984.

[2] T. Kohonen, *Self-organizing maps*. Springer, 1997.

[3] C. M. Bishop, M. Svensén, and C. K. I. Williams, "Developpements of generative topographic mapping," *Neurocomputing*, vol. 21, pp. 203–224, 1998.

[4] M. Girolami, "Document representation based on generative multivariate bernoulli latent topics models," in *BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research*, U. of Cambridge, Ed., 2001, pp. 194–201.

[5] A. Kabán and M. Girolami, "A combined latent class and trait model for analysis and visualisation of discrete data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 859–872, 2001.

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

[7] T. Hofmann and J. Puzicha, "Statistical models for co-occurrence data," MIT, Tech. Rep. AIM-1625, 1998.

[8] T. Hofmann, "Probmap - a probabilistic approach for mapping large document collections." *Intell. Data Anal.*, vol. 4, no. 2, pp. 149–164, 2000.

[9] A. Kabán, E. Bingham, and T. Hirsimki, "Learning to read between the lines: The aspect bernoulli model," *SIAM International Conference on Data Mining (SIAM DM04)*, pp. 462–66, 2004.

[10] G. J. McLachlan and D. Peel, *Finite Mixture Models*. New York: John Wiley and Sons, 2000.

[11] A. Dempster, N. Laird, and D. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[12] D. J. C. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, 1992.

[13] P. McCullagh and J. Nelder, *Generalized linear models*. London: Chapman and Hall, 1983.

[14] D. Bohning, "Construction of reliable maximum likelihood algorithms with application to logistic and cox regression," *Handbook of Statistics*, vol. 9, pp. 409–422, 1993.

[15] L. K. Saul, T. Jaakkola, and M. I. Jordan, "Mean field theory for sigmoid belief networks," *Journal of Artificial Intelligence Research*, vol. 4, pp. 61–76, 1996.

[16] O. Elemento, "Initialisation, convergence, et validation de cartes topologiques de kohonen (in french)," Master's thesis, Rapport de DEA (INRIA, Yves Lechevallier), 1999.

[17] I. Jolliffe, *Principal Component Analysis*. Springer Verlag, 2002.

[18] J. P. Benzecri, *Correspondence Analysis Handbook*. New-York : Dekker, 1992.

[19] S. Deerwester, S. T. Dumais, G. W. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[20] J. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computers*, vol. 5, no. 18C, pp. 401–409, may 1969.

[21] C. M. Bishop, *Neural Networks for Pattern Recognition*. Clarendon Press, 1995.

[22] R. Priam, "Méthodes de carte auto organisatrice par mélange de lois contraintes. application 1'exploration dans les tableaux de contingence textuels (in french)," Ph.D. dissertation, Université de Rennes 1, Octobre 2003.

[23] J. Zhang, "The mean field theory in EM procedures for markov random fields," *IEEE Transactions on Signal Processing*, vol. 10, no. 40, pp. 2570–2583, 1992.

[24] G. Celeux, F. Forbes, and N. Peyrard, "Em procedures using mean field-like approximations for markov model-based image segmentation," *Pattern Recognition*, vol. 36, pp. 131–144, 2003.

[25] C. Ambroise and G. Govaert, "Convergence of an em-type algorithm for spatial clustering," *Pattern Recogn. Lett.*, vol. 19, no. 10, pp. 919–927, 1998.

[26] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Trans. on Neural Networks*, vol. 11, no. 3, pp. 586–600, May 2000.

[27] I. Lerman, *Classification et Analyse Ordinale des Données*. Dunod, Paris, 1981.

[28] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003)*, 2003, pp. 89–98.

[29] G. Govaert and M. Nadif, "Clustering with block mixture models," *Pattern Recognition*, vol. 36, no. 2, pp. 463–473, 2003.

[30] ——, "An EM algorithm for the block mixture model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 643–647, 2005.