Automatic Relevance Determination in Nonnegative Matrix Factorization

Vincent Y. F. Tan, Cédric Févotte

Automatic Relevance Determination in Nonnegative Matrix Factorization

Vincent Y. F. Tan
Lab. for Info. and Decision Systems,
Massachusetts Institute of Technology,
Cambridge, MA 02139
Email: vtan@mit.edu

Cédric Févotte
CNRS LTCI; TELECOM ParisTech
37-39 rue Dareau
75014 Paris, France
Email: fevotte@telecom-paristech.fr

***Abstract*—Nonnegative matrix factorization (NMF) has become a popular technique for data analysis and dimensionality reduction. However, it is often assumed that the number of latent dimensions (or components) is given. In practice, one must choose a suitable value depending on the data and/or setting. In this paper, we address this important issue by using a Bayesian approach to estimate the latent dimensionality, or equivalently, select the model order. This is achieved via *automatic relevance determination* (ARD), a technique that has been employed in Bayesian PCA and sparse Bayesian learning. We show via experiments on synthetic data that our technique is able to recover the correct number of components, while it is also able to recover an *effective* number of components from real datasets such as the MIT CBCL dataset.**

***Index Terms*—Nonnegative matrix factorization, Bayesian model selection, Automatic relevance determination.**

I. Introduction

Nonnegative matrix factorization (NMF) [7] is a widely used technique that is employed for non-subtractive, part-based representation of nonnegative data. There are numerous diverse applications of NMF including audio signal processing [5], image classification [6] and email surveillance [1]. Given a nonnegative data matrix $\mathbf{V} \in \mathbb{R}_+^{F\times N}$, NMF essentially involves finding two nonnegative matrices, namely $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K] \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} = [h_1^T, \ldots h_K^T]^T \in \mathbb{R}_+^{K\times N}$ such that

$$\mathbf{V} \approx \hat{\mathbf{V}} \triangleq \mathbf{W}\,\mathbf{H}. \tag{1}$$

K is usually chosen such that $FK + KN \ll FN$, hence reducing the data dimension. NMF is also usually performed by minimizing a pre-specified cost function $D(\mathbf{V}|\mathbf{WH})$ over $\mathbf{W}$ and $\mathbf{H}$.

In practice, however, it is often hard to "guess" the effective dimension of the latent subspace K_{eff} i.e. the effective number of columns of $\mathbf{W}$ (and the number of rows of $\mathbf{H}$). To ameliorate this problem, in this paper, we formulate a Bayesian approach to find the effective dimensionality K_{eff} via *automatic relevance determination* [8]. This model selection technique was also used by Bishop to determine K_{eff} in the context of Bayesian PCA [2] and by Tipping in sparse Bayesian learning [10]. It is important to discover K_{eff} because we would like a parsimonious (sparse) yet accurate representation of the nonnegative data. Sparsity here refers to the total number of coefficients $K(F+N) \ll FN$ required to encode the data, and not only to $\mathbf{H}$ as in the standard sparse linear regression setting where the dictionary $\mathbf{W}$ is known and fixed.

A. Main contributions

To solve this important model selection problem, we place priors, dependent on precision-like parameters $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_K] \in \mathbb{R}_+^K$, on both the columns of $\mathbf{W}$ and the rows of $\mathbf{H}$. The values of these hyperparameters, together with the values of $\mathbf{W}$ and $\mathbf{H}$ are inferred iteratively by maximizing the posterior of the parameters given the data. As a result of this optimization, a subset of the β_k's will be driven to a large upper bound [2], [8], corresponding to irrelevant components. More precisely, we seek to find $\mathbf{W}^*, \mathbf{H}^*, \boldsymbol{\beta}^*$ by optimizing the maximum a-posteriori (MAP) criterion:

$$\min_{\mathbf{W},\mathbf{H},\boldsymbol{\beta}} C_{\text{MAP}}(\mathbf{W},\mathbf{H},\boldsymbol{\beta}) \triangleq -\log p(\mathbf{W},\mathbf{H},\boldsymbol{\beta}|\mathbf{V}), \tag{2}$$

where, by Bayes rule, the posterior can be written as

$$\begin{aligned} -\log\, p(\mathbf{W},\mathbf{H},\boldsymbol{\beta}|\mathbf{V}) &\stackrel{\text{c}}{=} -\log p(\mathbf{V}|\mathbf{W},\mathbf{H}) \\ &\quad -\log p(\mathbf{W}|\boldsymbol{\beta}) - \log p(\mathbf{H}|\boldsymbol{\beta}) - \log p(\boldsymbol{\beta}). \end{aligned} \tag{3}$$

We use $\stackrel{\text{c}}{=}$ to denote equality up to a constant. For some statistical models, maximizing the log-likelihood term $\log p(\mathbf{V}|\mathbf{W},\mathbf{H})$ in (3) over $(\mathbf{W},\mathbf{H})$ is equivalent to minimizing a specific cost function (or measure of fit) $D(\mathbf{V}|\mathbf{WH})$ over $(\mathbf{W},\mathbf{H})$. For instance, maximum likelihood estimation of $\mathbf{W}$ and $\mathbf{H}$ in Poisson noise model corresponds to minimizing the (generalized) Kullback-Leibler (KL) divergence $D_{KL}(\mathbf{V}|\mathbf{WH}) = D_{KL}(\mathbf{V}|\hat{\mathbf{V}})$, where

$$D_{KL}(\mathbf{V}|\hat{\mathbf{V}}) \triangleq \sum_{fn} v_{fn}\log\left(\frac{v_{fn}}{\hat{v}_{fn}}\right) - v_{fn} + \hat{v}_{fn}. \tag{4}$$

We focus on the KL-divergence as our cost function in this paper because it has been extensively used in NMF and is free of noise parameters. The model order selection technique we present in this paper naturally extends to other cost functions such as the Euclidean distance or Itakura-Saito divergence, that may also be mapped to log-likelihood functions [5]. The Euclidean or Itakura-Saito cost functions, however, require the specification or inference of noise parameters. We show in Section IV that our method is able to recover the correct number of components for synthetic datasets, including the Swimmer dataset [4], and an effective number of component for the MIT CBCL face dataset.

B. Related Work

There is generally little literature about model order selection in NMF. Variational Bayesian methods have been

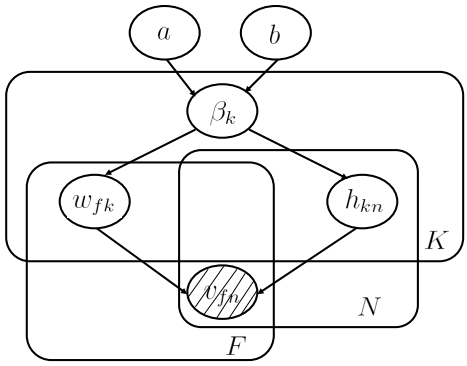

Figure 1: A directed graphical model that describes our NMF statistical model. The plate notation is used here. The plate (or rectangle) is used to group random variables that repeat. The number of replicates is shown on the bottom right corner. The shaded variables $\{v_{fn}\}$ are the observations.

considered for this task [3], [11] but this type of approach is fairly computationally demanding and require successive evaluation of the evidence (or marginal likelihood) for each considered value of K. Instead, our method automatically finds the optimal model order K_{eff} given a large initial value of K and is computationally less involving.

The methodology presented in this work bears resemblance with existing multiplicative NMF algorithms with a sparseness constraint on the whole matrix $\mathbf{H}$ (typically, adding a $\|\mathbf{H}\|_1$ or $\|\mathbf{H}\|_0$ regularization term to the cost function $D(\mathbf{V}\|\mathbf{WH})$), with which is was noted that the sparseness constraint may turn off excess components [9]. Here, we do not impose sparseness *per se* on either $\mathbf{W}$ or $\mathbf{H}$, but rather assign a relevance weight β_k to each rank-1 matrix $\mathbf{w}_k h_k$, which is learnt from the data.

II. Bayesian NMF model

In this section, we describe our statistical model. We will describe in detail each of the terms in (3), namely the log-likelihood term $\log p(\mathbf{V}|\mathbf{W},\mathbf{H})$ and the log-prior terms $\log p(\mathbf{W}|\boldsymbol{\beta})$, $\log p(\mathbf{H}|\boldsymbol{\beta})$ and $\log p(\boldsymbol{\beta})$.

A. Likelihood model

We assume that the log-likelihood of a single element of the matrix $\mathbf{V}$, denoted $p(v_{fn}|\hat{v}_{fn})$, is given by $\mathcal{P}(v_{fn}|\hat{v}_{fn})$, where $\mathcal{P}(x|\lambda) = e^{-\lambda}\lambda^x/\Gamma(x+1)$ is the Poisson probability density function with rate λ. Thus, $\mathbb{E}[v_{fn}|\hat{v}_{fn}] = \hat{v}_{fn}$. The log-likelihood can then be written as:

$$-\log p(\mathbf{V}|\mathbf{W},\mathbf{H}) = D_{KL}(\mathbf{V}|\mathbf{WH}) + \sum_{fn} v_{fn}(1-\log v_{fn}) + \log\Gamma(v_{fn}+1). \quad (5)$$

Thus, maximizing the log-likelihood (based on the Poisson noise model) is equivalent to minimizing the KL-divergence since other terms in (5) are independent of $\mathbf{W}$ and $\mathbf{H}$.

B. Prior model on $\mathbf{W}$ and $\mathbf{H}$

In Bayesian PCA [2], each column k of $\mathbf{W}$ (resp. row k of $\mathbf{H}$) is given a normal prior with precision parameter β_k (resp. 1). Thanks to the simplicity of the statistical model (multivariate Gaussian observations with Gaussian parameter priors), $\mathbf{H}$ can be easily integrated out of the likelihood, and optimization can be done over $p(\mathbf{W},\boldsymbol{\beta}|\mathbf{V})$. This marginalization step is not easily done in NMF. However, similar in spirit to Bayesian PCA, we define independent half-normal priors over each column k of $\mathbf{W}$ and row k of $\mathbf{H}$, and the priors are tied together through a *single*, common precision parameter β_k.[1] More precisely, we set

$$p(w_{fk}|\beta_k) = \mathcal{HN}(w_{fk}|0,\beta_k^{-1}), \quad (6)$$
$$p(h_{kn}|\beta_k) = \mathcal{HN}(h_{kn}|0,\beta_k^{-1}), \quad (7)$$

where

$$\mathcal{HN}(x|0,\beta^{-1}) = \sqrt{\frac{2}{\pi}}\beta^{-1/2}\exp\left(-\frac{1}{2}\beta x^2\right) \quad (8)$$

is the half-normal probability density function (defined for $x \geq 0$) parameterized by the precision (inverse variance) β.[2] From (6) and (7), we see that the minus log-priors can be written as:

$$-\log p(\mathbf{W}|\boldsymbol{\beta}) \stackrel{c}{=} \sum_k \sum_f \frac{1}{2}\beta_k w_{fk}^2 - \frac{F}{2}\log\beta_k, \quad (9)$$

$$-\log p(\mathbf{H}|\boldsymbol{\beta}) \stackrel{c}{=} \sum_k \sum_n \frac{1}{2}\beta_k h_{kn}^2 - \frac{N}{2}\log\beta_k. \quad (10)$$

As a result of inference, a subset of the β_k's will be driven to a large upper bound, with the corresponding columns of $\mathbf{W}$ and rows of $\mathbf{H}$ driven to small values, by (8). The effective dimensionality can be deduced from the distribution of the β_k's, from which, we have found in practice, two clusters clearly emerge : a group of values in same order of magnitude corresponding to relevant components and a group of similar values of much higher magnitude corresponding to irrelevant components. We may then define K_{eff} as

$$K_{\text{eff}} \stackrel{\Delta}{=} |\{\beta_k \ : \ \beta_k < L_k - \epsilon\}|, \quad (11)$$

where L_k is the upper bound dependent on the prior's parameters and $\epsilon > 0$ is a user-defined small constant. We will specify a precise value for L_k in terms F, N and the parameters of the prior on β_k in Section III.

This Bayesian approach based on ARD automatically determines the number of components in $\mathbf{V}$, thus elegantly finding a solution for the odd dichotomy between finding the best model for the training data to the model and avoiding overfitting. Furthermore, the size of β_k provides an approximate measure of its relevance. If β_k is large (variance $1/\beta_k$ is small), then this component is not very relevant for the modeling of the data. Conversely, if β_k is small (variance $1/\beta_k$ is large), the corresponding component contains a significant amount of energy. We term β_k the *relevance weight* associated to component k.

We remark that it is also possible to let $\mathbf{W}$ be dependent of $\boldsymbol{\beta}$ and let $\mathbf{H}$ be *independent* of $\boldsymbol{\beta}$. In this setup, we place a prior on the columns of $\mathbf{W}$ only. Doing this is akin to Bayesian PCA [2]. However, in our experiments, we found convincing evidence that tying $\mathbf{w}_k$ and h_k together with the same prior structure and precision parameter is better for recovery of the number of components. It is also possible

[1] Note that this prior is not overconstraining the scales, because of the scale indeterminacy between $\mathbf{w}_k$ and h_k.

[2] If X is a zero mean normal random variable with variance β^{-1}, then $|X|$ is a half-normal random variable with precision β.

to put exponential distributions, (dependent on the precision parameters $\beta_1, \ldots, \beta_K$) instead half-normals, on either $\mathbf{W}$ or $\mathbf{H}$ (or both). Indeed, the inference method described next would only need minor changes to accomodate other priors.

C. Prior model on $\boldsymbol{\beta}$

Finally, each precision parameter β_k is given a Gamma distribution, which is conjugate to the half-normal density function (8). Thus,

$$p(\beta_k|a_k, b_k) = \frac{b_k^{a_k}}{\Gamma(a_k)} \beta_k^{a_k - 1} \exp(-\beta_k b_k), \quad \beta_k \geq 0, \quad (12)$$

and the (minus) log-prior of $\boldsymbol{\beta}$ is given as

$$-\log p(\boldsymbol{\beta}) \stackrel{c}{=} \sum_k b_k \beta_k - (a_k - 1) \log \beta_k. \quad (13)$$

In our experiments, the hyperparameters a_k and b_k (also known as shape and scale parameters) are fixed to a pre-specified common value, say a and b. Thus, all the upper bounds $L_k = L$, some constant dependent on a, b.

D. Overall Model and Cost Function

All the terms in in the MAP criterion (3) have been defined. The dependences between the variables can be represented by a directed graphical model (Bayesian network) as shown in Fig. 1. The observations $\{v_{fn}\}$ are contained in the matrix $\mathbf{V}$ and we would like to infer $\mathbf{W}$, $\mathbf{H}$ and $\boldsymbol{\beta}$. The number of β_k's that is less than the upper bound $L_k - \epsilon$ is the effective dimensionality of the data.

By combining equations (5), (9), (10) and (12), the overall MAP objective to be minimized can be written as

$$\begin{aligned} &-\log p(\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}|\mathbf{V}) \stackrel{c}{=} D_{KL}(\mathbf{V}|\mathbf{W}, \mathbf{H}) \\ &+ \frac{1}{2} \sum_k \left[\left(\sum_f w_{fk}^2 + \sum_n h_{kn}^2 + 2b_k \right) \beta_k \right. \\ &\left. - (F + N - 2(a_k - 1)) \log \beta_k \right]. \end{aligned} \quad (14)$$

From this expression, we see a tradeoff involving the size of the β_k's. Since $F + N$ is typically larger than $2(a_k - 1)$, a larger β_k's will result in a smaller (more negative) last term. However, the second to last term in the sum forces some of the β_k's to remain small since it is a linear function of β_k.

III. Inference

In [7], the authors proposed efficient coordinate descent algorithms with suitable step sizes which are then turned into multiplicative update rules. This simple yet effective procedure has the advantage of maintaining nonnegativity of the inferred matrices $\mathbf{W}$ and $\mathbf{H}$. The same procedure can be applied to our model to infer $\mathbf{W}$, $\mathbf{H}$ as well as $\boldsymbol{\beta}$. We first find the gradient of the MAP criterion C_{MAP} with respect to $\mathbf{W}$, $\mathbf{H}$ and β_k.

$$\nabla_{\mathbf{W}} C_{\text{MAP}}(\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}) = \left(\frac{\mathbf{WH} - \mathbf{V}}{\mathbf{WH}} \right) \mathbf{H}^T + \mathbf{W} \text{diag}\left(\boldsymbol{\beta}\right), \quad (15)$$

$$\nabla_{\mathbf{H}} C_{\text{MAP}}(\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}) = \mathbf{W}^T \left(\frac{\mathbf{WH} - \mathbf{V}}{\mathbf{WH}} \right) + \text{diag}\left(\boldsymbol{\beta}\right) \mathbf{H}, \quad (16)$$

Algorithm 1 Automatic relevance determination for NMF with the KL-divergence cost

Input : Nonnegative data (observation) matrix $\mathbf{V}$, fixed hyperparameters a_k, b_k.
Output : Nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$ such that $\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{W}\,\mathbf{H}$, a nonnegative vector $\boldsymbol{\beta}$ and K_{eff}.
Initialize $\mathbf{W}$ and $\mathbf{H}$ to nonnegative values.
for $i = 1 : n_{iter}$ **do**
  $\mathbf{H} \leftarrow \frac{\mathbf{H}}{\mathbf{W}^T \mathbf{1}_{F \times N} + \text{diag}(\boldsymbol{\beta})\mathbf{H}} \cdot \left[\mathbf{W}^T \left(\frac{\mathbf{V}}{\mathbf{WH}}\right)\right]$
  $\mathbf{W} \leftarrow \frac{\mathbf{W}}{\mathbf{1}_{F \times N}\mathbf{H}^T + \mathbf{W}\text{diag}(\boldsymbol{\beta})} \cdot \left[\left(\frac{\mathbf{V}}{\mathbf{WH}}\right)\mathbf{H}^T\right]$
  $\boldsymbol{\beta} \leftarrow \frac{F+N+2(\mathbf{a}-1)}{\mathbf{1}_{1\times F}(\mathbf{W}\cdot\mathbf{W}) + (\mathbf{H}\cdot\mathbf{H})\mathbf{1}_{N\times 1} + 2\mathbf{b}}$
end for
Compute K_{eff} as in (11).

$$\nabla_{\beta_k} C_{\text{MAP}}(\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}) = \frac{1}{2} \sum_f w_{fk}^2 + \frac{1}{2} \sum_n h_{kn}^2 + b_k - \left(\frac{1}{2}(N + F) + a_k - 1 \right) \frac{1}{\beta_k}. \quad (17)$$

where we use the notation $\frac{\mathbf{A}}{\mathbf{B}}$ to denote entry-wise division of $\mathbf{A}$ by $\mathbf{B}$. The multiplicative coordinate descent algorithm is simple; we simply update each parameter θ of $\mathbf{W}$ and $\mathbf{H}$ by multiplying its current value with the ratio of the positive $[\cdot]_+$ to negative $[\cdot]_-$ part of the derivative of C_{MAP} (in (15) and (16)) with respect to θ i.e.,

$$\theta \leftarrow \theta \, \frac{[\nabla_\theta C_{\text{MAP}}(\theta)]_+}{[\nabla_\theta C_{\text{MAP}}(\theta)]_-}. \quad (18)$$

The precision parameters β_k are updated by zeroing (17) i.e.,

$$\beta_k \leftarrow \frac{\frac{1}{2}(F + N) + a_k - 1}{\frac{1}{2}(\sum_f w_{fk}^2 + \sum_n h_{kn}^2) + b_k}. \quad (19)$$

The overall algorithm is summarized in Algorithm 1. We observed in practice the monotonicity of the MAP criterion in (2) under this algorithm.

We now address the issue of choosing $L_k = L$ in (11). From (19), we can easily derive an upper bound on β_k, which is dependent only on a and b, since we set $a_k = a$ and $b_k = b$:

$$\beta_k \leq \frac{F + N + 2(a - 1)}{2b}. \quad (20)$$

The bound is attained when $w_{fk} = h_{kn} = 0$ for all f, n, k i.e., the columns and rows of $\mathbf{W}$ and $\mathbf{H}$ are exactly zero. Thus, a good choice of L is precisely the right-hand-side of (20). Hence, we define

$$L \triangleq \frac{F + N + 2(a - 1)}{b}, \quad (21)$$

and the constant L will enable us to determine K_{eff} from the set of β_k's from (11).

It is worth noting that other inference algorithms (on graphs) such as variational Bayes or Gibbs sampling may be employed to find the posterior distributions of the parameters of interest. In this paper, we are only interested in point estimates of $\mathbf{W}, \mathbf{H}$ and $\boldsymbol{\beta}$, hence the efficient multiplicative update rules suffice.

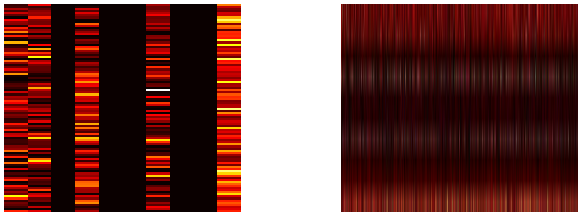

Figure 2: The $\mathbf{W}$ and $\mathbf{H}$ matrices showing that $K_{\text{eff}} = 5$ columns are retained. The other columns of $\mathbf{W}$ and rows of $\mathbf{H}$ are set to 0.

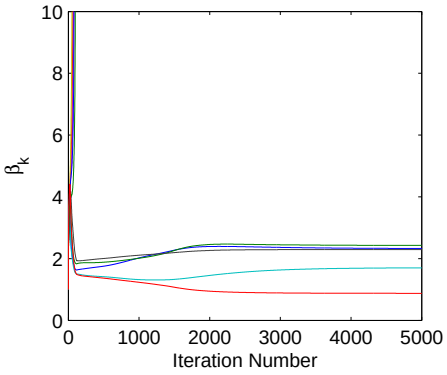

Figure 3: Convergence of $\boldsymbol{\beta}$. Note that $K - K_{\text{eff}} = 5$ components diverge to a large value.

IV. Numerical Results

In this section, we describe some experiments on synthetic and real data that demonstrate the efficacy of our algorithm for model order selection in NMF.

A. Synthetic Data

Our synthetic data was generated in the following fashion: $\mathbf{W}$ and $\mathbf{H}$ are 100×10 and 10×1000 matrices respectively. Each element in the first 5 columns of $\mathbf{W}$ (and the first 5 rows of $\mathbf{H}$) are drawn uniformly from a half normal distribution $\mathcal{HN}(0, 10)$ while the remaining 5 columns (remaining 5 rows of $\mathbf{H}$) are drawn from $\mathcal{HN}(0, 1)$. Thus the effective dimensionality is $K_{\text{eff}} = 5$. The 100×1000 data matrix $\mathbf{V}$ was formed by multiplying $\mathbf{W}$ and $\mathbf{H}$. The hyperparameters were set to $a = b = 1$, though in practice we found that the algorithm is not very sensitive to the choice of hyperparameters for the synthetic data. In Figs 2, we see that the algorithm correctly recovers the correct dimensionality of the data even though $\text{Rank}(\mathbf{V}) = K = 10$ with probability 1. In Fig 3, we observe that $K - K_{\text{eff}} = 5$ β_k's diverge to a large value, which means that 5 columns of $\mathbf{W}$ (and 5 rows of $\mathbf{H}$) are irrelevant for the purpose of modeling the $N = 1000$ data samples. We repeated the same procedure by fixing $K = 10$ and varied K_{eff} from 1 to 5. The algorithm correctly recovers K_{eff} in each case.

B. Swimmer Dataset

We also performed some experiments on another synthetic dataset, namely the well-known swimmer dataset [4], which depicts a figure with four moving parts (limbs), each able to exhibit four articulations (different positions). Thus, there are a total of $N = 4^4 = 256$ images each of size $F = 32 \times 32$. Sample images are shown in Fig. 4. The objective here is to discover that the dataset contains 16 basis elements, corresponding to the 16 limb positions. Thus, $K_{\text{eff}} = 16$. We set $K = 50$. In this experiment, we also found that we could retrieve the correct number of components, if b in the Gamma prior in (12) is larger than some threshold value. Keeping a fixed at 2, we plotted the *regularization path* in Fig. 5. The regularization path shows how K_{eff} changes as b in the Gamma prior in (12) changes, while being insensitive to the precise value of a. The components, which are the limbs, are plotted in Fig. 6 for $b = 18$. We see that each of the 16 limb positions is correctly recovered, together with the torso. All the relevance weights corresponding to the relevant components are equal, by equivalence of all the "parts" represented in the dataset ($\beta_k = 2.57$ for all $k = 1, \ldots, K_{\text{eff}}$). This experiment further validates our technique of automatically discovering the number of latent components in the nonnegative data.

Figure 4: Sample images from the Swimmer dataset.

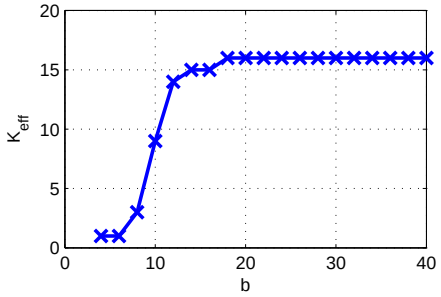

Figure 5: Regularization Path for Swimmer Dataset. The effective number of components K_{eff} is plotted against b, a parameter of the Gamma prior distribution in (12).

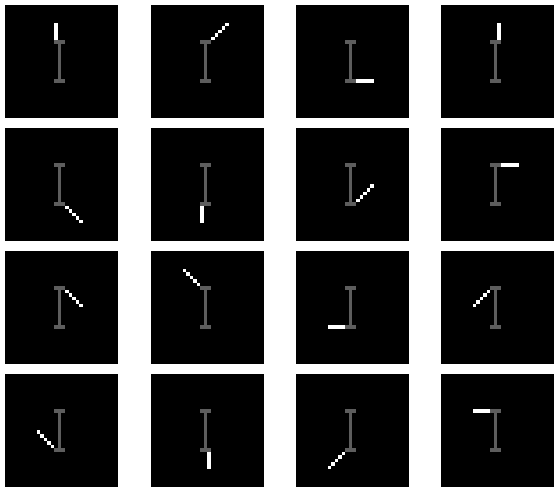

Figure 6: Basis images obtained by performing NMF on the Swimmer dataset. The 16 limb positions of the swimmer dataset are correctly recovered if $b \geq 18$. All the relevance weights $\{\beta_k\}$ are equal by symmetry.

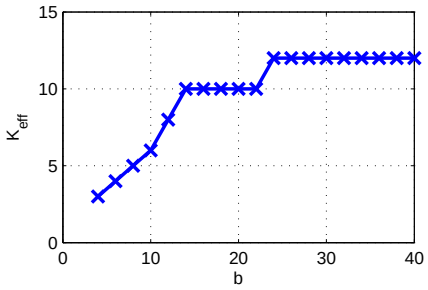

Figure 7: Regularization Path for MIT CBCL Dataset.

C. MIT CBCL Faces Dataset

Finally, we applied our model order selection procedure to the CBCL face dataset, containing $F = 19 \times 19$ images, preprocessed as in [7], i.e., the greyscale intensities of each image were linearly scaled so that the pixel mean and standard deviation were equal to 0.25, and then clipped to the range [0,1]. The training set contains $N = 2429$ images. Keeping a (the shape parameter in the Gamma prior in (12)) fixed at 2, we plot the regularization path in Fig. 7, i.e., the effective number of components K_{eff} against the scale parameter b. This figure shows that if b is above some threshold value, the number of components K_{eff} stabilizes at a constant value ($K_{\text{eff}} = 12$). K was set to 49 initially. For instance, if we set $b = 25$, we observe that 12 basis images (columns of $\mathbf{W}$) are recovered. These are shown in Fig. 8, together with the corresponding the relevance weights β_k's. They correspond to several parts of the faces : we roughly obtain left, right, top and bottom halves, background artifacts and eyes/eyebrows, mouth and nose parts.

Our adaptation of ARD to NMF allows us to automatically discover a latent dimensionality of this dataset and the recovered basis images make intuitive sense. However they are not as local as the one obtained with the standard NMF algorithm with K set to 49, as in the original experiment by Lee & Seung [7]. Other prior structures (such as the exponential distribution instead of the half-normal, or even sharper priors) may lead to more local features. However the results are subjective in every case and should instead be evaluated on well-defined tasks such as classification.

V. Conclusions

We have presented a Bayesian approach that performs model order selection for NMF by borrowing ideas from ARD. By placing appropriate prior distributions on the elements on $\mathbf{W}$ and $\mathbf{H}$, we are able to identify those components that are 'relevant' for modeling the data. The efficacy of our method was adequately demonstrated in our experiments where we are able to recover the latent dimensionality of synthetic and real data.

Indeed, our method is not limited to the use of the generalized KL-divergence; we could, in fact, adapt the updates in a very straightforward manner for the Euclidean or Itakura-Saito costs and also use other types of priors. Furthermore, our method can be extended for the purpose of nonnegative tensor decomposition (NTD). Our preliminary results show that it is possible to recover the correct number of components from a synthetically generated 3-way tensor

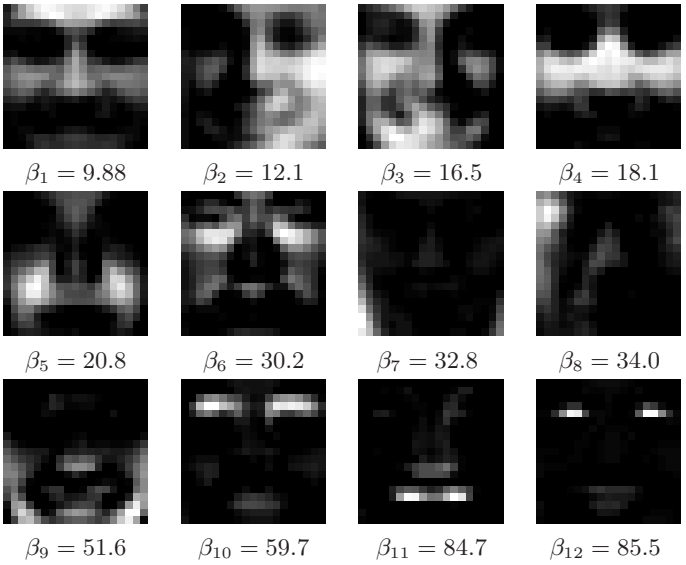

Figure 8: Basis images (reordered) and their corresponding relevance weights $\{\beta_k\}$ obtained by performing Bayesian model order selection using ARD for NMF on the MIT CBCL dataset. The images correspond to the parts of the face including left, right, top and bottom halves, the eyes, the mouth, background and nose.

formed from the parallel factor analysis (PARAFAC), also named canonical decomposition (CANDECOMP), model. We hope to further extend the theory and experimentation for tensor decomposition in a longer paper.

Acknowledgements

The work of V. Y. F. Tan was supported by A*STAR, Singapore, and by a MURI funded through ARO Grant W911NF-06-1-0076. C. Févotte acknowledges support from the ANR - SARAH project (Standardisation du Remastering Audio Haute-définition).

References

[1] M. W. Berry and M. Browne. Email Surveillance Using Non-negative Matrix Factorization. *Computational and Mathematical Organization Theory*, 11(3):249–264, 2005.

[2] C. M. Bishop. Bayesian PCA. In *Advances in Neural Information Processing Systems*, pages 382–388, 1999.

[3] A. T. Cemgil. Bayesian Inference for Nonnegative Matrix Factorisation Models. Technical Report CUED/F-INFENG/TR.609, Cambridge University Engineering Department, Jul 2008.

[4] D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems*, 2003.

[5] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence with application to music analysis. *Neural Computation*, 21(3), 2009.

[6] D. Guillamet, B. Schiele, and J. Vitri. Analyzing non-negative matrix factorization for image classification. In *Internat. Conf. Pattern Recognition*, 2002.

[7] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999.

[8] D. J. C. Mackay. Probable networks and plausible predictions – a review of practical Bayesian models for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995.

[9] M. Mørup, K. H. Madsen, and L. K. Hansen. Approximate L0 constrained non-negative matrix and tensor factorization. In *Proc. IEEE International Symposium on Circuits and Systems (ISCAS'08)*, 2008.

[10] M. E. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211 – 244, Sep 2001.

[11] O. Winther and K. B. Petersen. Bayesian independent component analysis: Variational methods and non-negative decompositions. *Digital Signal Processing*, 17(5):858–872, Sep. 2007.