# TUNING PRUNING IN SPARSE NON-NEGATIVE MATRIX FACTORIZATION

*Morten Mørup and Lars Kai Hansen*

Intelligent Signal Processing, DTU Informatics.
Richard Petersens Plads bld 321, 2800, Lyngby, Denmark
phone: + (45) 45253900, fax: + (45) 45872599, email: {mm,lkh}@imm.dtu.dk
web: www.mortenmorup.dk

## ABSTRACT

Non-negative matrix factorization (NMF) has become a popular tool for exploratory analysis due to its part based easy interpretable representation. Sparseness is commonly invoked in NMF (SNMF) by regularizing by the $l_1 - norm$ both to alleviate the non-uniqueness of the NMF representation as well as promote sparse (i.e. part based) representations. While sparseness can prune excess components thereby potentially also establish the number of components it is an open problem what constitutes the adequate degree of sparseness, i.e. how to tune the pruning. In a hierarchical Bayesian framework SNMF corresponds to imposing an exponential prior while the regularization strength can be expressed in terms of the hyper-parameters of these priors. Thus, within the Bayesian modelling framework Automatic Relevance Determination (ARD) can learn these pruning strengths from data. We demonstrate on three benchmark NMF data how the proposed ARD framework can be used to tune the pruning thereby also estimate the NMF model order.

## 1. INTRODUCTION

Non-negative matrix factorization NMF has become an important tool for unsupervised, exploratory data analysis. NMF decomposes a non-negative matrix $\mathbf{V}^{N\times M}$ into a positive low rank approximation (p-rank) given by

$$\mathbf{V}^{N\times M} = \mathbf{W}^{N\times K}\mathbf{H}^{K\times M} + \mathbf{E}^{N\times M}. \quad (1)$$

where $\mathbf{V} \geq \mathbf{0}, \mathbf{W} \geq \mathbf{0}, \mathbf{H} \geq \mathbf{0}$ (where $\geq$ denotes element-wise greater than zero). The decomposition is particularly useful because it results in easy interpretable part based representations [18]. Non-negative decompositions is also named positive matrix factorization [25] but was popularized by Lee and Seung due to a simple algorithmic procedure based on multiplicative updates [17]. The decomposition has proven useful for a wide range of data where non-negativity is a natural constraint. These encompass data for text-mining based on counts, image data, biomedical data and spectral data.

Unfortunately, the NMF decomposition is not in general unique [9, 15] neither is the decomposition in general guaranteed to admit sparse/part based representation. To overcome these limitations of NMF sparseness has been imposed on one of the modes forming the sparse-NMF (SNMF) [11, 10, 12] this has been achieved by regularizing using the $l_1$-norm. The benefit of the $l_1$-norm being that it is the closest convex proxy to minimizing the cardinality, i.e. number of non-zero elements. Optimizing for sparse representation is related to the classic rotation criteria such as VARIMAX [13] and maximum Likelihood independent component analysis (ICA) based on sparse priors [23]. However, rather than rotating an estimated solution, the estimation process is directly posed as a tradeoff between simplicity of the representation and fitting the data. Thus, a sparse representation is strongly related to the principle of parsimony, i.e., among all possible accounts the simplest is considered the best. If no formal prior information is given parsimony can be considered a reasonable guiding principle to avoid overfitting, see also [23] and references therein. Despite the great attention given to the NMF/SNMF decompositions over the past years two important open problems remain: *(1) What is the adequate number of components K in the* NMF *decomposition? (2) What degree of sparseness should be imposed in* SNMF*?* We will address these two problems using a standard approach in Bayesian inference referred to as Automatic Relevance Determination (ARD) [20, 2, 26]. Traditionally, ARD has been based on Gaussian priors yielding a ridge regression type of selection. In line with SNMF we will derive an ARD approach based on the exponential prior corresponding to regularizing by the $l_1$-norm. Thus, finding the right number of components as well as the right degree of sparseness can be turned into the single problem of tuning the pruning in SNMF [1]. In [24] we have demonstrated how ARD can be adapted to tensor factorization based on the PARAFAC and Tucker models. Presently we will demonstrate the use of ARD for the SNMF-problem. To investigate the impact of choice of NMF-objective we will derive the ARD approach for the two most used NMF objectives namely the least squares (LS) and Kullback-Leibler (KL) divergence [17] which in the Bayesian framework correspond to assuming Gaussian and Poisson distributed noise.

The paper is structured as follows. We will first state the SNMF problem in a Bayesian framework and use the ARD approach to tune the pruning in SNMF. We will next briefly investigate the performance of various commonly used maximum a posteriori (MAP) estimation algorithms for least squares NMF - to establish viable approaches for estimating the ARD-SNMF model parameters. We finally evaluate the proposed framework on three benchmark datasets.

## 2. METHODS

### 2.1 Tuning pruning by ARD

Automatic Relevance Determination (ARD) is a hierarchical Bayesian approach widely used for model selection [20, 26, 2]. In ARD hyperparameters explicitly represent the relevance of different features by defining the range of variation for these features, usually by modeling the width of

[1] We note that in [11] sparseness was controlled by a user defined sparseness degree of the components, however, this does not answer the question what constitutes the right degree of sparsity.

a zero-mean Gaussian prior imposed on the model parameters. If the width becomes zero, the corresponding feature cannot have any effect on the predictions. Hence, ARD optimizes these hyperparameters to discover which features are relevant. Traditionally ARD has been based on Gaussian priors as these are conjugate priors for the unconstrained least squares (LS) estimation problem. However, Gaussian priors truncated to the positive orthant (to form non-negative priors) are no-longer conjugate. Furthermore, they do not promote sparse representation within the active component since the resulting $l_2$-regularization penalizes elements by their squares and as such penalizes large values relatively more than small values. The exponential prior on the other hand is known to admit sparse representation as it corresponds to a $l_1$-regularization – the closest convex proxy to minimizing for the number of non-zero elements in the model [8]. Since $l_1$-regularization favors sparse representation and has been the preferred method of regularization for the SNMF problem in the literature [12, 11, 10] we will presently consider the $l_1$ regularized SNMF problem given by the exponential prior due to its sparsity promoting behavior. As we further want to turn off excess components thereby optimizing for $K$ we impose the following component-wise exponential prior on $\mathbf{H}$, i.e.

$$P(\mathbf{H}|\beta) = \prod_k \beta_k^M e^{-\beta_k \sum_m \mathbf{H}_{k,m}}. \quad (2)$$

We will derive the ARD approach both for least squares (LS-NMF) and KL-divergence (KL-NMF) optimization but we note that the approach readily generalizes to other NMF objectives such as Bregman and Alpha divergence [6, 4]. Our parameterizations will be in line with the formulation of the SNMF approach, however, we note that the ARD framework for NMF was recently also proposed in [30] for the KL-divergence based on a different parameterizations of the prior. In a Bayesian framework, the least squares (LS) objective and KL-divergence (KL), i.e.

$$\text{LS-NMF} \quad : \quad \frac{1}{2}\sum_{n,m}(\mathbf{V}_{n,m} - (\mathbf{WH})_{n,m})^2,$$

$$\text{KL-NMF} \quad : \quad \sum_{n,m}\mathbf{V}_{n,m}\log\frac{\mathbf{V}_{n,m}}{(\mathbf{WH})_{n,m}} + (\mathbf{WH})_{n,m} - \mathbf{V}_{n,m},$$

correspond to minimizing the negative log-likelihood assuming the entries in $\mathbf{V}$ are independent, identically distributed (i.i.d.) with Gaussian and Poisson noise respectively, i.e.
$P_{\text{LS}}(\mathbf{V}|\mathbf{W},\mathbf{H},\sigma^2) = \prod_{n,m}\frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{(\mathbf{V}_{n,m}-(\mathbf{WH})_{n,m})^2}{2\sigma^2}] \Rightarrow$
$\log P_{\text{LS}}(\mathbf{V}|\mathbf{W},\mathbf{H},\sigma^2) = const - NM\log\sigma - \frac{\|\mathbf{V}-\mathbf{WH}\|_F^2}{2\sigma^2}.$
$P_{\text{KL}}(\mathbf{V}|\mathbf{W},\mathbf{H}) = \prod_{n,m}\frac{e^{-(\mathbf{WH})_{n,m}}(\mathbf{WH})_{n,m}^{\mathbf{V}_{n,m}}}{\mathbf{V}_{n,m}!} \Rightarrow$
$\log P_{\text{KL}}(\mathbf{V}|\mathbf{W},\mathbf{H}) = const - (\mathbf{WH})_{n,m} + \mathbf{V}_{n,m}\log(\mathbf{WH})_{n,m}.$
As there is an inherent scale ambiguity between $\mathbf{W}$ and $\mathbf{H}$ we will as proposed in [10] for SNMF require that $\|\mathbf{w}_k\|_F = 1$ which corresponds to the improper (i.e. un-normalizable) prior

$$P(\mathbf{W}) \propto \prod_k \delta(\|\mathbf{w}_k\|_F - 1) \quad (3)$$

In a Bayesian framework priors on the hyper-parameters $\beta$ could also be imposed, however, we will for simplicity impose non-informative priors on $\beta$. Using Bayes' theorem the log posterior likelihood can now be written for LS-SNMF and KL-SNMF as

$$\begin{aligned}\log P_{\text{LS}}(\mathbf{W},\mathbf{H}|\mathbf{V},\sigma^2,\beta) &\propto \log\left(P_{\text{LS}}(\mathbf{V}|\mathbf{W},\mathbf{H},\sigma^2)P(\mathbf{W})P(\mathbf{H}|\beta)\right)\\ &= -\tfrac{\|\mathbf{V}-\mathbf{WH}\|_F^2}{2\sigma^2} - \textstyle\sum_k \beta_k \sum_m \mathbf{H}_{k,m}\\ &\quad -NM\log\sigma + M\textstyle\sum_k\log\beta_k + const.\end{aligned}$$

$$\begin{aligned}\log P_{\text{KL}}(\mathbf{W},\mathbf{H}|\mathbf{V},\beta) &\propto \log\left(P_{\text{KL}}(\mathbf{V}|\mathbf{W},\mathbf{H})P(\mathbf{W})P(\mathbf{H}|\beta)\right)\\ &= \mathbf{V}_{n,m}\log(\mathbf{WH})_{n,m} - (\mathbf{WH})_{n,m} - \textstyle\sum_k\beta_k\sum_m\mathbf{H}_{k,m}\\ &\quad +M\textstyle\sum_k\log\beta_k + const.\end{aligned}$$

When constraining $\|\mathbf{w}_k\|_F = 1$.

Notice how the second lines in each of the two expressions above form the regular SNMF objectives while the normalization terms in the likelihood functions are given in the third lines. It is due to these normalization terms that we can learn the degree of regularization from data, i.e. tune the pruning parameter $\beta_k$. Differentiating the log likelihood with respect to $\sigma^2$, $\sigma^2$ can in theory also be learned from the data. However, estimating $\sigma^2$ from the data has a tendency of underestimating the value of $\sigma^2$ due to over-fitting, i.e. the models ability to fit noise. We therefore used the following more viable approach proposed in [24] to set $\sigma^2$ based on the assumption that the modelled signal ($\mathbf{WH}$) and noise ($\mathbf{E}$) are uncorrelated, $\sigma^2 = \|\mathbf{V}\|_F^2/(NM(1+10^{\text{SNR}/10}))$, where SNR is a user defined signal to noise ratio. In all the experiments we used a fixed value of SNR = 0dB assuming the same degree of signal as noise in the data. In [24] the sensitivity of this parameter to the obtained decomposition was investigated and it was found that the parameter had little impact for conservative choices of SNR. In Algorithm 1 the proposed ARD-SNMF approach is outlined. Notice, contrary to LS optimization the hyperparameter $\sigma^2$ in KL is absent (for brevity this correspond to $\sigma^2 = 1$). $\beta_k > 10^9\frac{\sqrt{\|\mathbf{V}\|_F^2}}{NM}$ is a threshold defining when components are removed while $\mathbf{W}_{\setminus k}$ denotes $\mathbf{W}$ with the $k$ component removed. The update of $\beta$ follows by solving $\frac{\partial\log P}{\partial\beta} = \mathbf{0}$. Since there is no analytic solution for the posterior moments of $\mathbf{H}$ we will base the estimation on maximum a posteriori (MAP) estimation (denoted by $\text{MAP}_{\text{LS/KL}}$). Alternatively more computationally involved sampling approaches can be invoked [28]. To solve the MAP parameter estimation problem the various algorithms for regular NMF can be used. We will therefore in the next section briefly review some common approaches for solving the NMF problem.

**Algorithm 1** ARD-SNMF

set $K$ large enough to encompass all potential models,
SNMF-LS: $\sigma^2 = \|\mathbf{V}\|_F^2/(NM(1+10^{\text{SNR}/10}))$
set $\beta = \mathbf{0}$ and initialize by random $\mathbf{W}$ and $\mathbf{H}$.
**repeat**
  $\mathbf{W} \leftarrow \text{MAP}_{\text{LS/KL}}(\mathbf{V},\mathbf{H},\mathbf{W})$
  $\mathbf{H} \leftarrow \text{MAP}_{\text{LS/KL}}(\mathbf{V},\mathbf{H},\mathbf{W},\sigma^2\beta)$
  $\beta_k = \frac{M}{\sum_m \mathbf{H}_{k,m}}$
  **if** $\beta_k > 10^9\frac{\sqrt{\|\mathbf{V}\|_F^2}}{NM}$ **Then** $\mathbf{W} = \mathbf{W}_{\setminus k}, \quad \mathbf{H} = \mathbf{H}_{\setminus k}$ **endif**
**until** convergence

### 2.2 Solving the MAP-estimation problem efficiently

Over the last couple of years several algorithms have been proposed for the NMF problem. Most methods employ an

alternating strategy where $\mathbf{W}$ is updated for fixed $\mathbf{H}$ and $\mathbf{H}$ for fixed $\mathbf{W}$. Unfortunately, most comparison in the literature between algorithms have been based on performance pr. iteration, see also [14, 27] despite that performance relative to time is most often of main interest. We will therefore compare some of the most widely used LS-NMF approaches described in the following paragraph in terms of performance over time (as measured by Matlab cpu-time usage).

For brevity we will denote the gradient of the NMF objective function with respect to each alternating subproblem by $\mathbf{G}$ while $\theta$ will denote the parameter under consideration. Roughly the alternating NMF methods can be split into approaches that solve these subproblems exact and approximate. Classic exact methods to solve for the least squares NMF problem include Lawson and Hanson's active set procedure [16]. Here the unconstrained solution of the active set is calculated and optimally projected back to the positive orthant, an important computational improvement for the NMF problem has been to reuse the established active set from the previous iteration [3]. We will denote this method ACTSET. Approximate methods for solving each alternating subproblem include the widely used multiplicative updates (MU) proposed in [17], projected gradient PG [19] as well as what we denote (NAIVEALS) which corresponds to solving the least squares solution and truncating negative values to zero. In MU the gradient is decomposed into the non-negative quantities $\mathbf{G}^+$ and $\mathbf{G}^-$ such that $\mathbf{G} = \mathbf{G}^+ - \mathbf{G}^-$ and the parameters are updated according to $\theta \leftarrow \theta(\frac{\mathbf{G}^-}{\mathbf{G}^+})$. In [27] it was noted that the convergence of these MU could be improved by tuning a step-parameter $\mu$ exponentiating the gradient $\theta \leftarrow \theta(\frac{\mathbf{G}^-}{\mathbf{G}^+})^\mu$ forming MU that are adaptive (MUA). In PG the NMF problem is solved using the regular gradient descent updates truncating negative values to zero, i.e. $\theta \leftarrow \theta - \mu G, \quad \theta(\theta < 0) = 0$ where $\mu$ is tuned by line-search. As noted in [19] the main computational burden when calculating the gradient is the computation of $\mathbf{VH}^\top$ respectively $\mathbf{W}^\top\mathbf{V}$ having complexity $\mathcal{O}(NMK)$ whereas updating the gradient and evaluating the LS-objective (when running several updates of one mode while keeping the other mode fixed) has computational complexity $\mathcal{O}(NK^2)$ and $\mathcal{O}(MK^2)$ for each of the two alternating updates respectively. As a result, since $K \ll \min(N,M)$ it is computationally efficient to take several gradient steps at each alternating step. Thus, we also included an exact gradient search scheme such that each alternating step was terminated when the relative change in the LS-objective was less than $10^{-6}$ or 25 gradient steps had progressed forming the EMUA and EPG.

From figure 1 it can be seen that MU, MUA and EMUA despite being widely used updating strategies for NMF suffer extremely poor convergence whereas the just as simple PG and EPG strategy convergence to the solutions found by the exact ACTSET procedure[2] while relying solely on gradient information, in particular EPG seems to exhibit quadratic convergence as ACTSET at a much lower computational cost.

## 3. RESULTS

We evaluated the proposed ARD-SNMF approach on three benchmark dataset; the USPS handwritten image data set [5] containing 7,291 images in a 256 dimensional space, the CBCL Face Database ♯1 (MIT Center For Biological and Computation Learning `http://www.ai.mit.edu/projects/cbclcontaining`) containing 2,429 facial images in 361 dimensional space and the inter trial phase coherence ITPC of wavelet transformed EEG data measured across 14 subjects through left and right hand stimulation described in [21, 22] available from `http://www.erpwavelab.org`[3]. To investigate the choice of objective we compared the performance assuming Gaussian noise with the performance using Poisson noise, i.e. ARD-SNMF-LS and ARD-SNMF-KL respectively. For comparison we included the evaluation of model selection criteria based on Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) which have traditionally been used as simple approximations to the expectation of the negative log likelihood and the model evidence respectively [1, 29]: $\text{AIC} = -2\log L + Q$ and $\text{BIC} = -2\log L + Q\log(NM)$ where $L$ is the likelihood of the model, $Q = K(N+M)$ is the number of parameters in the model, and $NM$ the number of data points. Thus, the criteria define a tradeoff between reduction in reconstruction error and complexity of the model. Notice that BIC tends to penalize model complexity more heavily than AIC, hence, gives a more conservative estimate of what is considered the best model. It is an open problem how many components adequately describe the three datasets. However, based on visual inspection three components have been found to adequately describe the ITPC-data. We will impose sparseness on the second mode such that as much information as possible is coded in $\mathbf{W}$, i.e. the feature images for the USPS and CBCL data and channel mode of the ITPC data. To avoid the impact of local minima the best model, i.e. model with highest $logP$ value of 10 runs is given, at initialization we set $K = 100$ to encompass all potential models.

In figure 2 it is seen that ARD-SNMF-LS and ARD-SNMF-KL both extract a 10 component model for the USPS data, a 2 and 85 component model for the CBCL data and a 3 and 0 component model for the ITPC data. In figure 3 model selection as indicated by AIC and BIC is given.

## 4. DISCUSSION

While the analysis over the 10 runs of both ARD-SNMF-LS and ARD-SNMF-KL were highly consistent across the 10 random initialization indicating that the approach was not very prone to local minima the model order estimated was very dependent on the likelihood functions imposed, presently Gaussian vs. Poisson noise. While both models indicated that 10 components adequately modelled the USPS data the results of the two approaches significantly differed for the CBCL and ITPC data. While the ARD-SNMF-LS method extracted the previously reported three components of the ITPC data [21, 22] both BIC and AIC failed to indicate these components. Furthermore, the difference in the number of extracted components for ARD-SNMF-LS and ARD-SNMF-KL emphasize that the choice of noise model greatly impacts the components found. Thus, despite the choice of ARD prior as reported in [24] has limited impact on the components found, it is seen from the results of figure 2 that the choice of noise model greatly impacts the number of compents extracted.

Compared to information criteria such as AIC and BIC that has to evaluate all potential models the benefit of the

[2] MU can in fact be arbitrarily slow since for $\mathbf{G}^+ = \mathbf{Q} + \mathbf{R}$, $\mathbf{G}^- = \mathbf{Q}$ the gradient $\mathbf{G} = \mathbf{G}^+ - \mathbf{G}^- = \mathbf{R}$ whereas $\frac{\mathbf{G}^-}{\mathbf{G}^+} \to 1$ for $\mathbf{Q} \to \infty$.

[3] baseline coherence of $\sqrt{\frac{1}{\sharp trials}} = \sqrt{\frac{1}{360}}$ was subtracted the data.

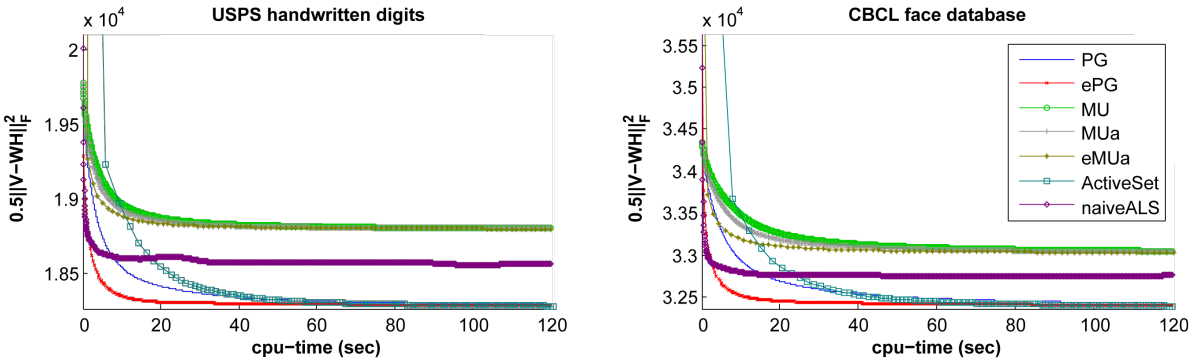


Figure 1: Comparison of various LS-NMF-optimization approaches by $\frac{1}{2}\|\mathbf{V}-\mathbf{WH}\|_F^2$ vs Matlab cpu-time for a computational budget of 120 seconds based on the USPS handwritten data (left panel) and CBCL face data (right panel). MU: multiplicative updates of [17], MUA: adaptive multiplicative updates of [27], EMUA: MUA with up to 25 gradient steps at each alternating iteration, PG: projected gradient with one gradient step, EPG: projected gradient with up to 25 gradient steps at each alternating iteration as proposed in [19], ACTSET: The active set procedure given in [3], NAIVEALS: Least squares solution where negative values are truncated to zero. All methods used same initial solution and apart from the updating rules all methods were based on identical Matlab implementations. Clearly, the MU based approaches do not converge to the same quality of solution as the PG based approaches. For KL-NMF optimization similar results were obtained (not shown) when comparing the PG and MU based approaches. Notice, when starting the PG in the obtained MU solutions we were able to recover the optimal solution.

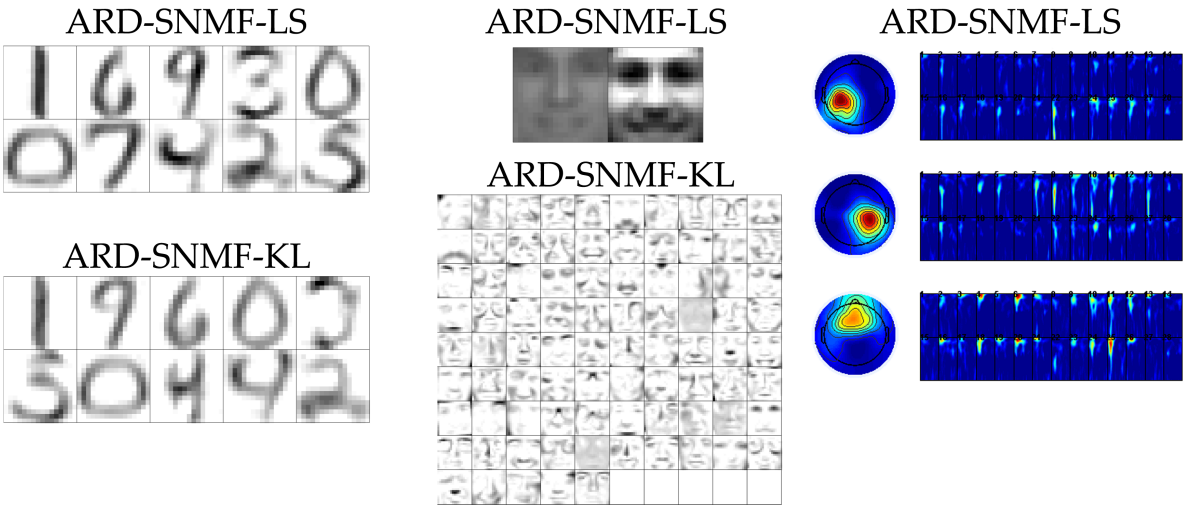


Figure 2: The estimated components by ARD-SNMF-LS (top panels) and ARD-SNMF-KL (bottom panels) for the USPS digit data, CBCL face data and ITPC EEG-data. For the digit data both methods extract 10 components whereas for the face data a 2 component and 85 component model have been estimated. For the ITPC data on the other hand the ARD-SNMF-LS extracts 3 components identifying left and right as well as frontal activation as reported in [22, 21] whereas the ARD-SNMF-KL has pruned all the components to zero.

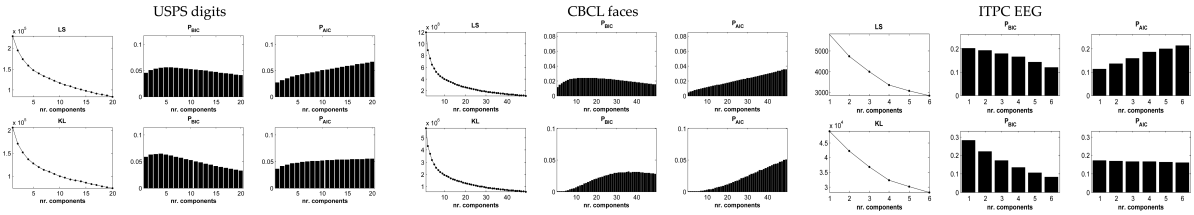


Figure 3: AIC and BIC analysis of the USPS digit data (left), CBCL face data (middle) and ITPC EEG-data (right) for ARD-SNMF-LS (top panels) and ARD-SNMF-KL (bottom panels). The AIC and BIC values have been turned into probabilities by $P_{\mathrm{BIC}_k} = \frac{e^{-\frac{\mathrm{BIC}_k - \min(\mathrm{BIC})}{NM}}}{\sum_k e^{-\frac{\mathrm{BIC}_k - \min(\mathrm{BIC})}{NM}}}$. Little consensus as to what constitutes the best model order is found between AIC and BIC.

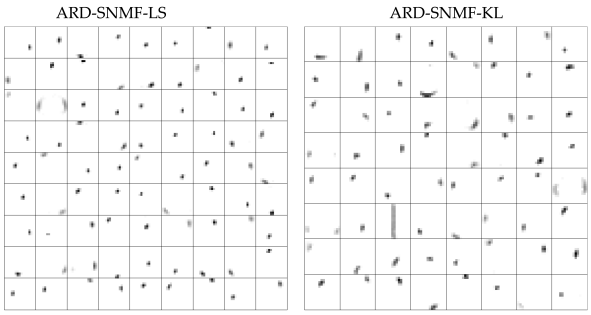


Figure 4: ARD-SNMF-LS and ARD-SNMF-KL analysis of the USPS handwritten data where sparseness is imposed on the pixel mode (i.e. the feature images) instead of on the mode coding each image. The two approaches have respectively extracted 81 and 64 very sparse image features.

present ARD approach is that the model order estimation comes at the cost of fitting one ordinary model while the approach seem to well extract the relevant activities in the data. However other parameterizations of the SNMF problem is conceivable. For the USPS and CBCL image data we imposed sparsity on $\mathbf{H}$ such that most information was coded in the feature images $\mathbf{W}$. Alternatively, sparseness could be imposed on the feature images instead by analyzing the transpose of $\mathbf{V}$ - such an analysis results instead in highly part based features as seen in figure 4 contrary to the analysis of figure 2 where the extracted features due to the sparsity on $\mathbf{H}$ seem to cluster the data into the different digit classes for the USPS data. We note that these two approaches to sparsity correspond well to the clustering aspects of NMF described in [7] and part based representation given in [18].

We presently considered the most computationally efficient but also simple model estimation based on MAP-estimates of the posterior likelihood functions. Within the Bayesian framework more involved approaches such as expectation propagation [26] as well as sampling methods [28] to estimate the distribution of the parameters can potentially improve the present ARD framework. Furthermore, within the hierarchical Baysiean framework priors on $\beta_k$ could also be imposed while other parameterizations of the priors are conceivable [30].

## Acknowledgement

This research was supported by the European Commission through the EU FP6 NEST Pathfinder grant PERCEPT (043261).

### REFERENCES

[1] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

[3] R. Bro and S. de Jong. A fast non-negativity-constrained least squares algorithm. *J. of Chemometrics*, 11(5):393–401, 1997.

[4] A. Cichocki, R. Zdunek, S. Choi, R. Plemmons, and S. Amari. Non-negative tensor factorization using alpha and beta divergences. *ICASSP*, 2007.

[5] Y. Le Cun, J. S. Denker, and S. A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*, pages 598–605, 1990.

[6] I. Dhillon and S. Sra. Generalized nonnegative matrix approximations with bregman divergences. In *Advances in Neural Information Processing Systems 18*, pages 283–290, 2005.

[7] C. Ding, X. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. *Proc. SIAM Internat. Conf. Data Min. (SDM'05)*, pages 606–610, 2005.

[8] D. Donoho. For most large underdetermined systems of linear equations the minimal $l^1$-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

[9] D. Donoho and V. Stodden. When does nonnegative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems 16*, 2003.

[10] J. Eggert and E. Körner. Sparse coding and nmf. In *Neural Networks*, volume 4, pages 2529–2533, 2004.

[11] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[12] P.O. Hoyer. Non-negative sparse coding. *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, pages 557–565, 2002.

[13] H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrica*, 23:187–200, 1958.

[14] D. Kim, S. Sra, and I. S. Dhillon. Fast projection-based methods for the least squares nonnegative matrix approximation problem. *Stat. Anal. Data Min.*, 1(1):38–51, 2008.

[15] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen. Theorems on positive data: On the uniqueness of nmf. *Computational Intelligence and Neuroscience*, 2008.

[16] C.L. Lawson and R.J. Hanson. *Solving Least Squares Problems*, volume 15 of *Classics in Applied Mathematics*. SIAM, Philadelphia, PA, 1995, 1974.

[17] D. D. Lee and H. S. Seung. Algorithms for nonnegative matrix factorization. In *NIPS*, pages 556–562, 2000.

[18] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–91, 1999.

[19] C.-J. Lin. Projected gradient methods for non-negative matrix factorization. *Neural Computation*, 19:2756–2779, 2007.

[20] D. J. C. Mackay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992.

[21] M. Mørup, L.K. Hansen, and S. M. Arnfred. Erpwavelab a toolbox for multi-channel analysis of time-frequency transformed event related potentials. *Journal of Neuroscience Methods*, 161(361-368), 2007.

[22] M. Mørup, L.K. Hansen, and S. M. Arnfred. Algorithms for sparse non-negative tucker. *Neural Computation*, 20(8):2112–2131, 2008

[23] M. Mørup. *Decomposition Methods for Unsupervised Learning*. PhD thesis, Technical University of Denmark, 2008.

[24] M. Mørup and L. K. Hansen. Automatic relevance determination for multi-way models. *accepted for publication, Journal of Chemometrics*, 22:1–12, 2009.

[25] P Paatero and U Tapper. Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[26] Y. Qi, T. P. Minka, R. W. Picard, and Z. Ghahramani. Predictive automatic relevance determination by expectation propagation. In *Proceedings of the twenty-first international conference on Machine learning*, page 85, New York, NY, USA, 2004. ACM.

[27] R. Salakhutdinov, S. Roweis, and Z. Ghahramani. On the convergence of bound optimization algorithms. In *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, pages 509–516, 2003.

[28] M. N. Schmidt, O. Winther, and L. K. Hansen. Bayesian non-negative matrix factorization. accepted for publication ICA 2009.

[29] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[30] V. Y. F. Tan and C. Fvotte. Automatic relevance determination in non-negative matrix factorization. *In Proc. Workshop on Signal Processing with Adaptative Sparse Structured Representations (SPARS09)*, 2009.