INFINITE NON-NEGATIVE MATRIX FACTORIZATION

Mikkel N. Schmidt and Morten Mørup

Technical University of Denmark, Richard Petersens Plads, DTU Bldg. 321, 2800 Kgs. Lyngby, Denmark

ABSTRACT

We propose the infinite non-negative matrix factorization (INMF) which assumes a potentially unbounded number of components in the Bayesian NMF model. We devise an inference scheme based on Gibbs sampling in conjunction with Metropolis-Hastings moves that admits cross-dimensional exploration of the posterior density. The approach can effectively establish the model order for NMF at a less computational cost than existing approaches such as thermodynamic integration and existing reversible jump Markov chain Monte Carlo sampling schemes. On synthetic and real data we demonstrate the success of (INMF).

1. INTRODUCTION

Non-negative matrix factorization NMF has become an important tool for unsupervised, exploratory data analysis due to its easily interpretable parts-based representation of data [16]. NMF decomposes a non-negative matrix $\boldsymbol{V} \in \mathbb{R}^{I \times J}$ into a positive low rank approximation (p-rank) given by

$$\underset{I\times J}{\boldsymbol{V}} = \underset{I\times D}{\boldsymbol{W}}\,\underset{D\times J}{\boldsymbol{H}} + \underset{I\times J}{\boldsymbol{E}}, \tag{1}$$

where the dimensions are indicated below each matrix, and $w_{id} \geq 0$, $h_{dj} \geq 0$ and e_{ij} is residual noise. Non-negative matrix factorization is also named positive matrix factorization [25] but was popularized by Lee and Seung due to a simple algorithmic model fitting procedure based on multiplicative updates [17]. The NMF decomposition has proven useful for a wide range of data where non-negativity is a natural constraint. Applications include text-mining based on word counts [16, 5], image analysis [16], neuro-informatics [22], bio-informatics [1], chemometrics [7], astronomy [26], and audio processing [31] to mention but a few. For a recent overview of NMF see also [3].

While NMF has found widespread use, an important open problem remains to efficiently determine the number of components D. Contrary to singular value decomposition (SVD) in which models with different number of components are nested, the components of the NMF decomposition change when D changes. Consequently, the interpretation of the decomposition relies on the number of extracted components and determining the model order is thus crucial in order to reliably interpret the components. Choosing the NMF model order amounts to estimating the posterior distribution of D (also denoted the marginal likelihood or evidence). Using Bayes' theorem, this is given by

$$p(D|\boldsymbol{V}) = \frac{p(\boldsymbol{V}|D)p(D)}{p(\boldsymbol{V})} \propto p(\boldsymbol{V}|D)p(D). \tag{2}$$

This, however, requires the computation of the quantity

$$p(\boldsymbol{V}|D) = \int p(\boldsymbol{V}|\boldsymbol{\Theta}, D)p(\boldsymbol{\Theta}|D)\mathrm{d}\boldsymbol{\Theta}, \tag{3}$$

where $\boldsymbol{\Theta}$ denotes the parameters of the NMF model. In general, this integral is analytically intractable and can be approximated using Markov chain Monte Carlo (MCMC).

Previous approaches to model order selection have considered a fixed range of model orders and carried out separate analyses for each D. This, however, might be a computationally wasteful approach: If the posterior in Eq. (2) is highly peaked it is not sensible to spend computational resources evaluating a possibly large range of very improbable model orders. In this paper we will make a first attempt to overcome these limitations by considering a non-parametric infinite non-negative matrix factorization (INMF) model where a potentially unbounded number of components can be considered without having to exhaustively evaluate all potential model orders in separate analyses.

1.1 MAP estimation of NMF

Traditionally, the NMF model has been fitted by various algorithms based on optimizing some error measure or computing maximum likelihood (ML) or maximum a posteriori (MAP) estimates of $\boldsymbol{W}$ and $\boldsymbol{H}$. In many of these approaches, the (non-convex) joint problem of estimating $\boldsymbol{W}$ and $\boldsymbol{H}$ is split into two (convex) sub-problems estimating $\boldsymbol{W}$ for fixed $\boldsymbol{H}$ and vice versa. Each sub-problem is commonly solved either by second order approaches such as the active set procedure [15, 13] or first order methods such as multiplicative updates [17] or projected gradient methods [18]. For an overview of estimation approaches see also [3, 12].

Several approaches to establish the model order based on MAP-parameter estimates have been proposed. The Bayesian information criteria (BIC) is an asymptotic expansion of the likelihood given in Eq. (3) such that the number of components are selected by minimizing the following quantity, $\mathrm{BIC} = -2\log L + K\log N$, where $L = p(\boldsymbol{V}|\boldsymbol{\Theta}^{\mathrm{MAP}}, D)$ is the likelihood, $\boldsymbol{\Theta}^{\mathrm{MAP}}$ is the MAP estimate of the parameters, K is the number of parameters, and N is the number of data points. For least squares estimation this reduces to $\mathrm{BIC} = N\log\frac{\mathrm{SSE}^{\mathrm{MAP}}}{N} + K\log N$ where $\mathrm{SSE}^{\mathrm{MAP}} = \|\boldsymbol{V} - \boldsymbol{W}^{\mathrm{MAP}}\boldsymbol{H}^{\mathrm{MAP}}\|_F^2$ is the residual sum of squared error of the MAP parameter estimates. Thus, the BIC criteria defines a tradeoff between model fit and complexity.

An alternative approach based on automatic relevance determination (ARD) has recently been applied in conjunction with MAP estimation of the NMF model [33, 21]. Here, priors on the model parameters are given hyper-parameters that represents the scale of each component by defining its range of variation. By optimizing these hyper-parameters, components can be removed if their scale goes below some threshold. This results in an estimate of the model order when the model is initialized with "too many" components.

Although MAP based approaches in general are very efficient they do not take parameter uncertainty into account, and as such only form an approximation to Eq. (2).

1.2 Bayesian NMF

To evaluate the integral in Eq. (3) Markov chain sampling approaches can be used to obtain a Monte Carlo estimate of the posterior distribution of the parameters, $p(\boldsymbol{\Theta}|\boldsymbol{V}, D)$. In [23, 32, 34] Gibbs sampling is used to obtain estimates of the joint posterior distribution of the NMF parameters $\boldsymbol{\Theta}$. In Gibbs sampling it is assumed that $\boldsymbol{\Theta}$ can be partitioned

into N groups, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$, such that it is possible to generate samples from the posterior conditional densities, $p(\boldsymbol{\theta}_n|\boldsymbol{\Theta}\backslash\boldsymbol{\theta}_n)$, for each of these groups. For the NMF model each element of a column of $\boldsymbol{W}$ and a row of $\boldsymbol{H}$ are conditionally independent such that the columns of $\boldsymbol{W}$ and rows of $\boldsymbol{H}$ can be sampled independently resulting in $N = 2D$ groups. In addition, parameters of the noise distribution and possible hyper-parameters must be sampled as well. Given some initial value of the parameters, each $\boldsymbol{\theta}_n$ is iteratively sampled while keeping all other parameters fixed. This procedure forms a homogeneous Markov chain that can be shown to sample from the full posterior distribution.

In our infinite NMF we use Gibbs sampling in conjunction with cross-dimensional Metropolis-Hasting moves. In the following, we consider an NMF model based on a Gaussian likelihood and rectified Gaussian priors,

$$v_{ij} \sim \mathcal{N}\Big(v_{ij}\Big|\sum_d w_{id}h_{dj}, \sigma^2\Big), \tag{4}$$

$$w_{id} \sim \mathcal{RG}(w_{id}|\mu_{id}, \tau_{id}^2), \tag{5}$$

$$h_{dj} \sim \mathcal{RG}(h_{dj}|m_{dj}, s_{dj}^2), \tag{6}$$

$$\sigma^2 \sim \mathcal{IG}(\sigma^2|\beta, \gamma), \tag{7}$$

where $\mathcal{N}(\cdot|\mu, \sigma^2)$ denotes the Gaussian density, $\mathcal{RG}(\cdot|\mu, \sigma^2) = \frac{2}{1+\mathrm{erf}(-\mu/\sigma)}\mathcal{N}(\mu, \sigma^2)\mathbf{1}(\cdot)$ denotes the rectified Gaussian density where $\mathbf{1}(\cdot)$ is a unit step function (see also [30]), and $\mathcal{IG}(\cdot|\beta, \gamma)$ denotes the inverse Gamma density. We note that the ideas presented here can be similarly applied to other NMF parameterizations such as [23, 32, 30, 34]. Our parameterization, $\boldsymbol{\Theta} = \{\boldsymbol{W}, \boldsymbol{H}, \sigma^2\}$, results in the following posterior conditional distributions required for the Gibbs sampler

$$w_{id}|\boldsymbol{V}, \boldsymbol{\Theta}\backslash w_{id} \sim \mathcal{RG}(w_{id}|\bar{\mu}_{id}, \bar{\tau}_{id}^2), \tag{8}$$

$$\bar{\tau}_{id}^2 = \big(\sum_j h_{dj}^2\sigma^{-2} + \tau_{id}^{-2}\big)^{-1}, \tag{9}$$

$$\bar{\mu}_{id} = \bar{\tau}_{id}^2\big(\tfrac{1}{\sigma^2}\sum_j h_{dj}(v_{ij} - \sum_{k\neq d} w_{ik}h_{kj}) + \tfrac{\mu_{id}}{\tau_{id}^2}\big), \tag{10}$$

$$h_{dj}|\boldsymbol{V}, \boldsymbol{\Theta}\backslash h_{dj} \sim \mathcal{RG}(h_{dj}|\bar{m}_{dj}, \bar{s}_{dj}^2), \tag{11}$$

$$\bar{s}_{dj}^2 = \big(\sum_i w_{id}^2\sigma^{-2} + s_{dj}^{-2}\big)^{-1}, \tag{12}$$

$$\bar{m}_{dj} = \bar{s}_{dj}^2\big(\tfrac{1}{\sigma^2}\sum_i w_{id}(v_{ij} - \sum_{k\neq d} h_{kj}w_{ik}) + \tfrac{m_{dj}}{s_{dj}^2}\big), \tag{13}$$

$$\sigma^2|\boldsymbol{V}, \boldsymbol{\Theta}\backslash\sigma^2 \sim \mathcal{IG}\big(\sigma^2|\beta + \tfrac{IJ}{2},\ \gamma + \tfrac{1}{2}\|\boldsymbol{V} - \boldsymbol{W}\boldsymbol{H}\|_F^2\big). \tag{14}$$

By iteratively sampling each parameter from their respective posterior conditional distributions, samples from the joint posterior distribution $p(\boldsymbol{\Theta}|\boldsymbol{V}, D)$ are obtained. Using this sample estimate, several approaches have been proposed to evaluate the NMF model order.

1.2.1 Chib's method

In [32] Chib's method [2] for model order estimation is applied to NMF. Here the marginal likelihood $p(\boldsymbol{V}|D)$ is obtained through the relation

$$p(\boldsymbol{V}|D) = \frac{p(\boldsymbol{V}|\bar{\boldsymbol{\Theta}}, D)p(\bar{\boldsymbol{\Theta}}|D)}{p(\bar{\boldsymbol{\Theta}}|\boldsymbol{V}, D)}, \tag{15}$$

where $\bar{\boldsymbol{\Theta}}$ is some high posterior density value of the parameters. The numerator can be directly evaluated while the denominator is approximated through N successive runs of the Gibbs sampler. As such the model requires the evaluation of all possible model orders in some set $\mathcal{D} = \{D_{\min}, \ldots, D_{\max}\}$, $\tilde{D} = |\mathcal{D}|$, resulting in a total of $N\tilde{D}$ posteriors densities to be estimated through Gibbs sampling.

1.2.2 Thermodynamic Integration

In [34] an NMF model order selection method based on thermodynamic integration [4] is proposed. Here, estimates of the marginal likelihood are derived through the use of power posteriors based on ideas from path sampling [6] from the prior to the posterior. A temperature parameter $t \in [0, 1]$ is imposed forming *power posterior*, $p_t(\boldsymbol{\Theta}|\boldsymbol{V}, D) = p(\boldsymbol{V}|\boldsymbol{\Theta}, D)^t p(\boldsymbol{\Theta})$, which is equal to the posterior for $t = 1$ and the prior for $t = 0$. The thermodynamic integral is then given by [4]

$$\log p(\boldsymbol{V}|D) = \int_0^1\int_{\boldsymbol{\Theta}} \log\,[p(\boldsymbol{V}|\boldsymbol{\Theta}, D)]p_t(\boldsymbol{\Theta}|\boldsymbol{V}, D)\mathrm{d}t\mathrm{d}\boldsymbol{\Theta}. \tag{16}$$

The integral over $\boldsymbol{\Theta}$ can be approximated by Gibbs sampling while the integral over t is carried out by considering a finite discretization of $t \in [0; 1]$. Thus, thermodynamic integration requires the estimation of the joint posterior for each model order $D \in \mathcal{D}$ and for each discretized temperature. For T temperatures and $\tilde{D}$ considered model orders, a total of $T\tilde{D}$ joint posteriors must then be estimated through Gibbs sampling.

1.2.3 RJMCMC

To overcome the high computational cost of Chib's method and thermodynamic integration, [34] proposes to use reversible jump Markov chain Monte Carlo sampling (RJMCMC) to obtain an estimate of Eq. (2). RJMCMC was first proposed by [8] and is a Metropolis-Hastings sampling approach that can perform cross-dimensional moves. Based on ideas from [19], [34] use independent proposal distributions based on approximations of the posterior for each model order. A drawback of this approach is thus that a separate Gibbs sampling run for each potential model order is required to obtain the proposal densities before the actual cross-dimensional sampling is used to estimate Eq (2). It is noted in [34] that

> "...it would be possible to add or remove some rows and columns of $[\boldsymbol{H}]$ and $[\boldsymbol{W}]$ and sample from some proposal distributions to jump between subspaces. However, this would not work as the samples would continually run out of mass of the extremely complex posterior distributions, and thus jumping from one subspace to another would never happen."

In the following, we present such an RJMCMC approach that jumps between subspaces based on adding or removing some rows and columns of $\boldsymbol{W}$ and $\boldsymbol{H}$, and demonstrate that by choosing good proposal densities cross-dimensional jumps are accepted with high probability. This allows for sampling all parameters as well as the model order jointly, eliminating the need for initially sampling from the posteriors of each possible model order. The inference scheme automatically infers the posterior distribution over the model order, and because the potential number of components is unbounded a priori we denote this method the infinite non-negative matrix factorization (INMF).

1.3 Existing infinite matrix factorization methods

Related to INMF, there exists a class of infinite matrix factorization approaches, including infinite binary matrix factorization (IBMF) [20], infinite sparse coding (ISC) and infinite independent component analysis (IICA) [14], that are based on the Indian buffet process (IBP) [9] which is a distribution over unbounded binary matrices. The IBMF model is given by $\boldsymbol{V} = \boldsymbol{U}\boldsymbol{Q}\boldsymbol{V}^\top + \boldsymbol{E}$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are binary matrices with a potentially infinite number of columns. Although attractive for its non-parametric representation, the binary constraints imposed on $\boldsymbol{U}$ and $\boldsymbol{V}$ make the model unable to account

well for general non-negative features as in NMF. The ISC and ICA models are given by $\boldsymbol{V} = \boldsymbol{A}(\boldsymbol{S} \odot \boldsymbol{Z}) + \boldsymbol{E}$, where $\boldsymbol{A}$ and $\boldsymbol{S}$ are general unbounded matrices, $\boldsymbol{Z}$ is an unbounded binary matrix, and $\odot$ denotes element-wise product. The model results in a sparse feature representation, where $\boldsymbol{A}$ are the extracted features, the binary matrix $\boldsymbol{Z}$ indicates which features are present for each data point, and $\boldsymbol{S}$ holds the real-valued coefficients of these features. In [14], a Gibbs sampling inference procedure is proposed, and with suitable prior densities a benefit of this model is that when estimating a given element of $\boldsymbol{Z}$ the corresponding element of $\boldsymbol{S}$ can be marginalized out analytically. By constraining $\boldsymbol{A}$ and $\boldsymbol{S}$ to be non-negative, the model corresponds to a sparse NMF representation [10]; however, the sparsity imposed through the binary activation pattern $\boldsymbol{Z}$ may not always comply well with the structure of the data if the assumption in NMF is true that all features are partially expressed to some degree in every data point.

2. INFINITE NMF

Rather than forming an infinite NMF model through the binary IBP representation we device a RJMCMC sampling procedure that can perform general cross-dimensional jumps efficiently. Cross-dimensional jumps from a model of order D with parameters $\boldsymbol{\Theta}$ to a model of order D^* with parameters $\boldsymbol{\Theta}^*$ is accepted with probability given by the reversible jump Metropolis-Hastings ratio

$$\min\left\{\frac{p(\boldsymbol{\Theta}^*, D^*|\boldsymbol{V})q(\boldsymbol{U}^*|\boldsymbol{\Theta}^*, D^*, \boldsymbol{V}, I^*)q(I|D)}{p(\boldsymbol{\Theta}, D|\boldsymbol{V})q(\boldsymbol{U}|\boldsymbol{\Theta}, D, \boldsymbol{V}, I)q(I^*|D^*)}, 1\right\}, \quad (17)$$

where $\boldsymbol{U}$ and $\boldsymbol{U}^*$ are auxiliary variables such that $n_{\boldsymbol{\Theta}} + n_{\boldsymbol{U}} = n_{\boldsymbol{\Theta}^*} + n_{\boldsymbol{U}^*}$ where $n_{\boldsymbol{\Theta}}$ denotes the number of elements in $\boldsymbol{\Theta}$. For ease of notation we have further included auxiliary variables I and I^*, which are index sets that point to a number of features, i.e., columns of $\boldsymbol{W}$ and the corresponding rows of $\boldsymbol{H}$. Given I, the cross-dimensional jump proposal is deterministic given by $(\boldsymbol{\Theta}^*, \boldsymbol{U}^*, I^*) = g(\boldsymbol{\Theta}, \boldsymbol{U}, I)$ where g is a bijective function with a Jacobian determinant of 1. (For that reason the Jacobian determinant term that usually occurs in the expression for the RJMCMC acceptance ratio is omitted.) The function g removes the features indexed by I from $\boldsymbol{\Theta}$ and places them into $\boldsymbol{U}^*$ and then appends the features in $\boldsymbol{U}$ to $\boldsymbol{\Theta}$ forming the new feature $\boldsymbol{\Theta}^*$, and I^* points to the indexes of the appended features. Finally, $q(I|D)$ denotes the probability of selecting a given feature index set for removal.

The crux for the RJMCMC procedure to be efficient is to achieve a reasonably high acceptance rate, which requires forming highly probable proposals $q(\boldsymbol{U}|\boldsymbol{\Theta}, D, \boldsymbol{V}, I)$. In the following, we consider two approaches for proposing cross-dimensional jumps: A birth-death procedure, which adds or removes one feature, and an split-merge procedure, which splits one feature into two or merges two to one. Both procedures are inspired by similar procedures for Dirichlet process mixtures [11].

The proposals are based on the idea of a launch state: Since $q(\boldsymbol{U}|\boldsymbol{\Theta}, D, \boldsymbol{V}, I)$ is allowed to depend on the existing features $\boldsymbol{\Theta}$, these can be used to deterministically compute an initial highly probable launch value, $\boldsymbol{U}^{\text{launch}}$, for the new features $\boldsymbol{U}$. As shown in [11], the computation of the launch state need not be deterministic: If the procedure is stochastic, it simply corresponds to a mixture transition, where a Markov chain transition is chosen randomly from a set of valid transitions. Here, we use the following procedure: We launch new features generated from the prior and refine them through t restricted Gibbs sweeps over the new features conditioned on the existing features less the removed features. Next, $q(\boldsymbol{U}|\boldsymbol{\Theta}, D, \boldsymbol{V}, I)$ can be defined as a random walk starting from $\boldsymbol{U}^{\text{launch}}$. Here, we use one final restricted Gibbs sweep, $q(\boldsymbol{U}|\boldsymbol{\Theta}, D, \boldsymbol{V}, I) = q(\boldsymbol{U}|\boldsymbol{\Theta}^{\text{launch}}, \boldsymbol{V})$, where $\boldsymbol{\Theta}^{\text{launch}} = g(\boldsymbol{\Theta}, \boldsymbol{U}^{\text{launch}}, I)$. The transition probability for the restricted Gibbs sweep can be computed as the product of the probabilities of each conditional parameter update given in Eq. (8–14).

2.1 Birth-death procedure

In the birth-death procedure we set the probability of generating a new component (birth) or removing a component (death) to be equal except if $D = 0$ where a death move has zero probability. As a result, we have the following contingency table for $q(I|D)$,

$q(I\|D)$	$D = 0$	$D > 0$	Move type
$I = \emptyset$	1	$\frac{1}{2}$	*Birth*
$I \in \{1, \ldots, D\}$	0	$\frac{1}{2D}$	*Death.*

(18)

Consequently, for a death move we randomly select a feature to remove among the D active features with probability $1/D$. For a birth move, we launch a new feature as explained above. The motivation behind the birth-death proposal is that it will allow the inclusion of extra features that model the residual error if needed, and conversely allow the deletion of unnecessary features.

2.2 Split-merge procedure

When inspecting the features of the NMF decompositions for different model orders it is often observed that including additional components has the effect that a previously observed component splits into two new different components. As observed by [16] NMF often results in parts-based representations, and adding additional components often results in existing parts being further atomized into smaller constituent parts. Based on this observation, we device a split-merge procedure that explicitly exploits this dynamic to generate highly probable proposal distributions for cross-dimensional jumps.

As for the birth-death approach we will assume that both a split and a merge step has equal probability except when $D = 1$ where a merge move has zero probability. As a result we have the following contingency table for $q(I|D)$,

$q(I\|D)$	$D = 1$	$D > 1$	Move type
$I \in \{1, \ldots, D\}$	1	$\frac{1}{2D}$	*Split*
$I = (i_1, i_2)$, $i_1 \neq i_2$, $i_1, i_2 \in \{1, \ldots, D\}$	0	$\frac{1}{2D(D-1)}$	*Merge.*

(19)

In a split move we randomly select an existing feature and remove it. We then launch two new features using the launch mechanism described above. In a merge move, we randomly select two different exiting features and remove them. Then we launch one new component with the slight modification of the procedure that the initial value of the new feature is taken as the average of the exiting features rather than generated from the prior before it is refined through t restricted Gibbs sweeps as before.

The birth-death procedure as well as the split-merge procedure are illustrated in Figure 1.

3. RESULTS

In the following we present simulations on toy examples as well as a real chemical shift imaging data set.

We generated four simple data sets by drawing from the prior: We generated two small 10×10 matrices with 3 components and two 100×100 matrices with 6 components at different noise levels (see Table 3). The priors were chosen

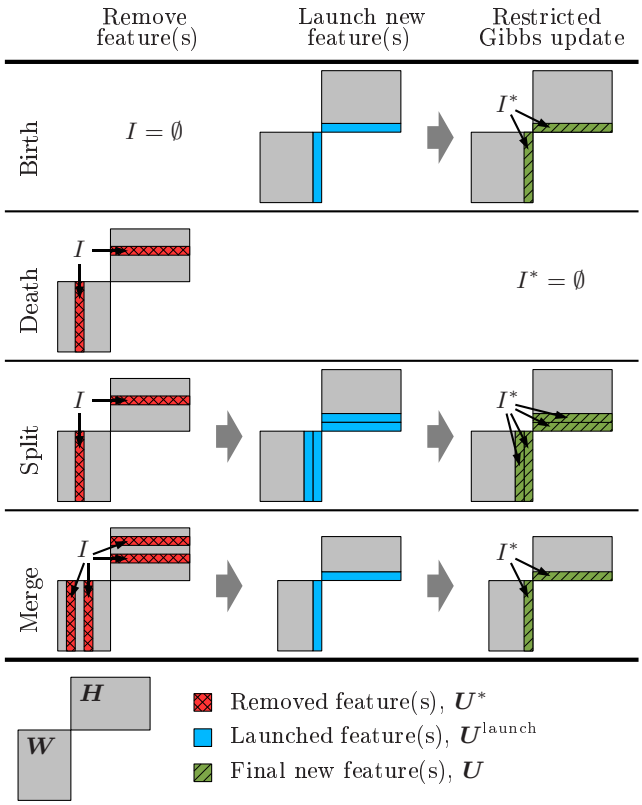

Figure 1: Illustration of the generated proposal densities for the birth-death (top) and split-merge (bottom) procedures. The features $\boldsymbol{U}^*$ indexed by I (red) are removed from $\boldsymbol{\Theta}$. New features are generated and refined through t restricted Gibbs sampling sweeps forming $\boldsymbol{U}^{launch}$ (blue). The final new features, $\boldsymbol{U}$, indexed by I^*, are generated through one final restricted Gibbs sweep, and the proposal density is computed by keeping track of all transition probabilities in the final Gibbs sweep.

as $\mu_{id} = m_{dj} = 0$, $\tau_{id} = s_{dj} = 1$, $\beta = 1$, $\gamma = \sigma^2$. We assumed a flat improper prior over the number of components, $p(D) \propto 1$. Initializing with $D = 0$, we then computed 10^6 posterior samples using the proposed inference procedure for INMF. In our experiments we interleaved one birth-death and split-merge proposal with five Gibbs sweeps, and used $t = 10$ restricted Gibbs sweeps to generate the launch states. The posterior probability of D for each data set is given in Figure 3. The results are as would be expected: For the two low noise data sets, A and C, the posterior is highly peaked around the correct number of components, whereas for the high noise data sets, B and D, the posterior is less peaked and skewed towards fewer components. In all examples the maximum posterior probability is at the correct model order. For comparison we implemented a naive version of the RJMCMC method in [34]. We computed a proposal density for each value of $D \in \{1, \ldots, 10\}$ based on fitting a rectified Gaussian to 10^6 posterior samples generated by Gibbs sampling from the model. Although we were able to obtain results similar to the ones obtained using INMF, we experienced that our naive implementation had severe difficulties mixing across dimensionalities. The NMF model has the inherent permutation ambiguity that any two features can be permuted resulting in the same posterior density. Thus, when the posterior samples reflect this, it should be taken into account in constructing a good proposal density, e.g.

	I	J	D	σ^2
A	10	10	3	1
B	10	10	3	10
C	100	100	6	1
D	100	100	6	10^3

Table 1: Toy example data sets.

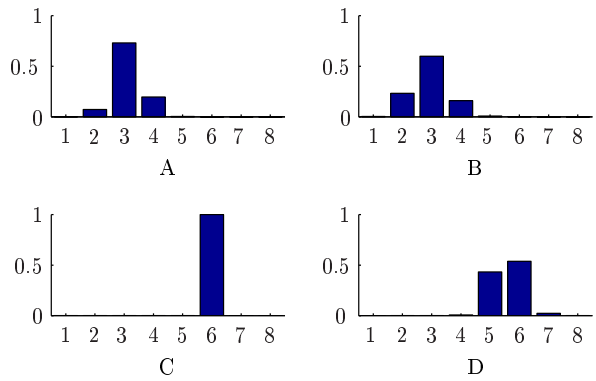

Figure 2: Results on toy data sets A-D: Posterior distribution of D. In the two low noise data sets (left) INMF finds the correct number of components with relatively high posterior probability. In the two high noise data sets (right) the posterior is less peaked and slightly skewed towards a smaller number of components.

by fitting a mixture model or by permuting the posterior samples.

Next, we analyzed a 369×256 chemical shift imaging data set [24] that has previously been analyzed using several NMF related methods [27, 28, 32, 29] and is known to contain two components. To match the noise level and scale of the data in line with [32], we chose the prior as $\mu_{id} = -10$, $m_{dj} = -10^6$, $\tau_{id} = 10$, $s_{dj} = 10^1 1$, $\beta = 1$, $\gamma = 10^8$. Initializing with $D = 0$, we computed 10^6 posterior samples using the INMF inference procedure similar as above. The estimated components matched the ones computed using other NMF related methods, and the posterior distribution of D had almost all of its mass at $D = 2$.

4. DISCUSSION

We proposed the infinite non-negative matrix factorization (INMF) model which has a potential unbounded number of features. We devised an efficient sampling scheme that were able to perform cross-dimensional jumps using Metropolis-Hastings moves. To avoid extreme low-probability proposals we derived high-probability configurations based on the calculation of an intermediate launch state as proposed for the Dirichlet process mixture in [11]. On synthetic and real data we demonstrated how the proposed approach was able to extract the underlying model order reliably at a lower computational cost than competing approaches such as Chib's method, thermodynamic integration, and the RJMCMC approach given in [34].

One might suspect that the presented procedure is nearly as computationally expensive as [34] due to the intermittent Gibbs sampling steps used to derive the launch states; however, these steps are only carried out on a small number of possible components, since the Markov chain predominantly explores the high probability region of the posterior. We do note, however, that the procedure is sensitive to the number of restricted Gibbs sweeps, t, and that we observed better

cross-dimensional mixing as t was increased, thus automatically selecting t remains as an important issue.

Although the method performed well we believe mixing can be further improved: If two components are (almost) identical and should be merged, there is a scale ambiguity between the components. Furthermore, there is an inherent scale ambiguity between $\boldsymbol{W}$ and $\boldsymbol{H}$ which is only partly resolved by the prior specification. We expect that extending the method to take these ambiguities into account, e.g. by including a scale constraint in the prior, might lead to better performance.

References

[1] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences (PNAS)*, 101(12):4164–4169, Mar 2004.

[2] S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, Dec 1995.

[3] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Wiley, 2009.

[4] N. Friel and A. N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society*, 70(3):589–607, 2008.

[5] E. Gaussier and C. Goutte. Relation between PLSA and NMF and implication. In *Research and Development in Information Retrieval, Proceedings of the International SIGIR Conference on*, pages 601–602, 2005.

[6] A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(5):163–185, 1998.

[7] C. Gobinet, E. Perrin, and R. Huez. Application of non-negative matrix factorization to fluorescence spectroscopy. *European Signal Processing Conference (EUSIPCO)*, pages 1095–1098, 2004.

[8] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrica*, 82(4):711–732, 1995.

[9] T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *Neural Information Processing Systems, Advances in (NIPS)*, pages 475–482, 2006.

[10] P. O. Hoyer. Non-negative sparse coding. In *Neural Networks for Signal Processing, IEEE Workshop on*, pages 557–565, 2002.

[11] S. Jain and R. M. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182, 2004.

[12] D. Kim, S. Sra, and I. S. Dhillon. Fast projection-based methods for the least squares nonnegative matrix approximation problem. *Statistical Analysis and Data Mining*, 1:38–51, 2008.

[13] H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *Matrix Analysis and Applications, SIAM Journal on*, 30(2):713–730, 2008.

[14] D. Knowles and Z. Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. In *Independent Component Analysis and Signal Separation, International Conference on (ICA)*, volume 4666 of *Lecture Notes in Computer Science Series (LNCS)*, pages 381–388. Springer, 2007.

[15] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, NJ, 1974.

[16] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[17] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Neural Information Processing Systems, Advances in (NIPS)*, pages 556–562, 2000.

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

[19] H. F. Lopes and M. West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14:41–67, 2004.

[20] E. Meeds, Z. Ghahramani, R. M. Neal, and S. T. Roweis. Modeling dyadic data with binary latent factors. In *Advances in Neural Information Processing Systems (NIPS)*, volume 19, pages 977–984, 2007.

[21] M. Mørup and L. K. Hansen. Tuning pruning in sparse non-negative matrix factorization. *European Signal Processing Conference (EUSIPCO)*, pages 1923–1927, 2009.

[22] 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.

[23] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret. Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling. *Signal Processing, IEEE Transactions on*, 54(11):4133–4145, Nov 2006.

[24] M. F. Ochs, R. S. Stoyanova, F. Arias-Mendoza, and T. R. Brown. A new method for spectral decomposition using a bilinear Bayesian approach. *Journal of Magnetic Resonance*, 137:161–176, 1999.

[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, Jun 1994.

[26] V. P. Pauca, J. Piper, and R. J. Plemmons. Nonnegative matrix factorization for spectral data analysis. *Linear Algebra and its Applications*, 416(1):29–47, Jul 2006.

[27] P. Sajda, S. Du, T. R. Brown, R. Stoyanova, D. C. Shungu, X. Mao, and L. C. Parra. Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain. *Medical Imaging, IEEE Transactions on*, 23(12):1453–1465, Dec 2004.

[28] P. Sajda, S. Du, and L. Parra. Recovery of constituent spectra using non-negative matrix factorization. In *Wavelets: Applications in Signal and Image Processing, Proceedings of SPIE*, volume 5207, pages 321–331, 2003.

[29] M. N. Schmidt and H. Laurberg. Nonnegative matrix factorization with Gaussian process priors. *Computational Intelligence and Neuroscience*, Feb 2008.

[30] M. N. Schmidt and S. Mohamed. Probabilistic non-negative tensor factorization using Markov chain Monte Carlo. *European Signal Processing Conference (EUSIPCO)*, pages 1918–1922, 2009.

[31] M. N. Schmidt and M. Mørup. Nonnegative matrix factor 2-d deconvolution for blind single channel source separation. In *Independent Component Analysis and Blind Signal Separation, International Conference on (ICA)*, volume 3889 of *Lecture Notes in Computer Science (LNCS)*, pages 700–707. Springer, Apr 2006.

[32] M. N. Schmidt, O. Winther, and L. K. Hansen. Bayesian non-negative matrix factorization. In *Independent Component Analysis and Signal Separation, International Conference on (ICA)*, volume 5441 of *Lecture Notes in Computer Science (LNCS)*, pages 540–547. Springer, 2009.

[33] V. Y. F. Tan and C. Fevotte. Automatic relevance determination in nonnegative matrix factorization. *Signal Processing with Adaptative Sparse Structured Representations, Workshop on (SPARS)*, 2009.

[34] M. Zhong and M. Girolami. Reversible jump MCMC for non-negative matrix factorization. In *Artificial Intelligence and Statistics, International Conference on (AISTATS)*, volume 5, pages 663–670, 2009.