# Reversible Jump MCMC for Non-Negative Matrix Factorization

**Mingjun Zhong**
mingjun@dcs.gla.ac.uk
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ, Scotland UK
http://www.dcs.gla.ac.uk/inference/

**Mark Girolami**
girolami@dcs.gla.ac.uk
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ, Scotland UK
http://www.dcs.gla.ac.uk/inference/

## Abstract

We present a fully Bayesian approach to Non-Negative Matrix Factorisation (NMF) by developing a Reversible Jump Markov Chain Monte Carlo (RJMCMC) method which provides full posteriors over the matrix components. In addition the NMF model selection issue is addressed, for the first time, as our RJMCMC procedure provides the posterior distribution over the matrix dimensions and therefore the number of components in the NMF model. A comparative analysis is provided with the Bayesian Information Criterion (BIC) and model selection employing estimates of the marginal likelihood. An illustrative synthetic example is provided using blind mixtures of images. This is then followed by a large scale study of the recovery of component spectra from multiplexed Raman readouts. The power and flexibility of the Bayesian methodology and the proposed RJMCMC procedure to objectively assess differing model structures and infer the corresponding plausible component spectra for this complex data is demonstrated convincingly.

## 1 INTRODUCTION

We consider the NMF problem (Paatero and Tapper, 1994; Lee and Seung, 2001) of representing a non-negative matrix $\mathbf{X}$ as a product of two non-negative matrices, formulated as,

$$\mathbf{X} = \mathbf{AS} + \mathbf{E}, \tag{1}$$

where $\mathbf{A} \in \mathcal{R}_+^{N\times M}$, $\mathbf{S} \in \mathcal{R}_+^{M\times T}$, and $\mathbf{E} \in \mathcal{R}^{N\times T}$ is the tolerance within each column and is assumed to follow a Normal distribution with zero mean and unknown diagonal covariance $\mathbf{\Lambda} = diag(\lambda_1, \cdots, \lambda_N)$. Each row of $\mathbf{S}$ designates one component and $M$ is the number of components (NoC). Our aim is to obtain the joint posterior for $M$, $\mathbf{A}$, and $\mathbf{S}$. It is well known that this decomposition is non-unique and hence not likelihood identifiable, we therefore invoke weak identifiability by the use of proper priors within the Bayesian framework.

Some NMF algorithms based on gradient methods, e.g., (Lee and Seung, 2001), could be directly employed to obtain *Maximum a Posteriori* estimates of $\mathbf{A}$ and $\mathbf{S}$ given an assumed NoC $M$. Indeed a Bayesian approach to NMF, using Metropolis sampling, is proposed in (Moussaoui *et al.*, 2006) although no attempt at model-order inference was made in that work. All of these methods implicitly assume that the NoC are known *a priori*. Estimating NoC is essentially the model selection problem and in this paper we consider several possible approaches. It is required to compute the posterior distribution of $M$ given observed data $\mathbf{X}$, which is proportional to the marginal likelihood $p(\mathbf{X}|M)$. However, the integral $p(\mathbf{X}|M) = \int p(\mathbf{X}, \mathbf{\Theta}|M)d\mathbf{\Theta}$ is analytically intractable for the current problem, where we denote $\mathbf{\Theta} = \{\mathbf{A}, \mathbf{S}, \mathbf{\Lambda}\}$. The BIC of (Schwarz, 1978) is a method of obtaining an asymptotic approximation of the marginal likelihood and could be simply employed with any number of standard NMF algorithms although as will be demonstrated in subsequent sections this approximation is not without its shortcomings. The use of the Thermodynamic Integral identity forms the basis for estimates of the log-marginal likelihood which have been shown to improve upon estimators using samples from the posterior density, see (Friel and Pettitt, 2008) and references therein. For example estimating the marginal likelihood using the harmonic mean identity can yield highly unstable estimators and ways to circumvent this problem have been proposed in (Raftery

Appearing in Proceedings of the $12^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. 

*et al.*, 2007). (Green, 1995) proposed the general RJMCMC methodology which is a generalized Metropolis-Hastings algorithm allowing trans-dimensional exploration of model and parameter space. Green's method considers $M$ as a random variable, and the posterior of all the model parameters, i.e., $P(M, \mathbf{\Theta}|\mathbf{X})$, is the invariant distribution. The outputs of the RJMCMC algorithm are then both the samples of $\mathbf{\Theta}$ and $M$.

## 2 A GIBBS SAMPLER FOR NMF

We firstly develop a Gibbs sampler for NMF which will be used for approximating the marginal likelihood and deriving an RJMCMC algorithm. The $t$th column of $\mathbf{X}$, $\mathbf{S}$ and $\mathbf{E}$ are represented as $\mathbf{x}_t$, $\mathbf{s}_t$ and $\epsilon_t$, respectively. The element of the $n$th row and $t$th column of $\mathbf{X}$ is $x_{nt}$, and with similar form the elements of $\mathbf{A}$ and $\mathbf{S}$ are $a_{nm}$ and $s_{mt}$ respectively. Given model (1) the data likelihood can be represented as $p(\mathbf{X}|\mathbf{A}, \mathbf{S}, \mathbf{\Lambda}) = \prod_{t=1}^{T} \mathcal{N}_{\mathbf{x}_t}(\mathbf{A}\mathbf{s}_t, \mathbf{\Lambda})$ where $\mathcal{N}_{\mathbf{y}}(\mu, \mathbf{\Sigma})$ represents the probability density function of a Normal random vector $\mathbf{y}$ with mean vector $\mu$ and covariance matrix $\mathbf{\Sigma}$. Note that all the vectors of this paper are column vectors for simplicity. In selecting our prior distributions for $s_{mt}, a_{nm}$, and $\lambda_n$ we note that a conjugate form for the variance of the likelihood takes the form of a Gamma density and to take account of possible sparseness a truncated exponential distribution is placed on each $a_{nm}$ (Hoyer, 2004). For the prior on each $s_{mt}$ it is assumed that the values will *a priori* be distributed uniformly in a certain range consistent with the observed data and so

$$p(s_{mt}|c_s, d_s) = \texttt{Unif}(c_s, d_s),$$
$$p(a_{nm}|\alpha_a, c_a, d_a) = \texttt{Expon}(\alpha_a)\mathbf{1}_{[c_a, d_a]}(a_{nm}),$$
$$p\left(\lambda_n^{-1}|\alpha_\lambda, \beta_\lambda\right) = \texttt{Gamma}(\alpha_\lambda, \beta_\lambda)$$

where $\mathbf{1}_A(\omega)$ denotes an indicator function, and the hyper-parameters $c_s, d_s, \alpha_a, c_a, d_a, \alpha_\lambda, \beta_\lambda$ are fixed in this paper. For the experiments we are considering, we observed that setting $\alpha_a = \alpha_\lambda = \beta_\lambda = 1$ was able to give acceptable results. We set $c_a = 0, d_a = 2$, and the $c_s > 0$ and $d_s > 0$ are set based on the observed data values. Now then we need to compute the full posterior $p(\mathbf{S}, \mathbf{A}, \mathbf{\Lambda}|\mathbf{X}, \gamma)$ for a fixed $M$ where $\gamma = \{\alpha_a, \alpha_\lambda, \beta_\lambda, c_s, d_s, c_a, d_a\}$. Details of derivations are shown in the APPENDIX. Note that the posteriors of the columns of $\mathbf{S}$ are independent as are the rows of $\mathbf{A}$. The conditional posterior of the $m$th row of $\mathbf{S}$ can be represented as follows,

$$p(\mathbf{s}_m|\mathbf{X}, \mathbf{A}, \mathbf{S}_{-m}) \propto \mathcal{N}_{\mathbf{s}_m}(\boldsymbol{\mu}_{s_m}, \mathbf{\Sigma}_{s_m}) \prod_{t=1}^{T} \mathbf{1}_{[c_s, d_s]}(s_{mt})$$

where $\mathbf{S}_{-m}$ represents those elements of $\mathbf{S}$ excluding $\mathbf{s}_m$, $\boldsymbol{\mu}_{s_m} = (\mu_{s_{m1}}, \cdots, \mu_{s_{mT}})^T$ and $\mathbf{\Sigma}_{s_m} = \sigma_{s_m}^2 \mathbf{I}$. As was noted before, $\mathbf{s}_m$ is a column vector, i.e., the transpose of the $m$th row of $\mathbf{S}$. Similarly, if we represent the $m$th column of $\mathbf{A}$ as $\mathbf{a}_m$, the posterior of $\mathbf{a}_m$ can therefore be represented as

$$p(\mathbf{a}_m|\mathbf{X}, \mathbf{A}_{-m}, \mathbf{S}) \propto \mathcal{N}_{\mathbf{a}_m}(\boldsymbol{\mu}_{a_m}, \mathbf{\Sigma}_{a_m}) \prod_{n=1}^{N} \mathbf{1}_{[c_a, d_a]}(a_{nm})$$

where $\mathbf{A}_{-m}$ represents those elements of $\mathbf{A}$ excluding $\mathbf{a}_m$, $\boldsymbol{\mu}_{a_m} = (\mu_{a_{1m}}, \cdots, \mu_{a_{Nm}})^T$ and $\mathbf{\Sigma}_{a_m} = diag(\sigma_{a_{1m}}^2, \cdots, \sigma_{a_{Nm}}^2)$. The conditional posterior of the reciprocal of the noise variance can be represented as a Gamma distribution with the following form,

$$p(\lambda_n^{-1}|\mathbf{x}_n, \mathbf{a}_n, \mathbf{S}, \alpha_\lambda, \beta_\lambda) = \texttt{Gamma}(\alpha_n, \beta_n)$$

where $\mathbf{x}_n$ represents the transpose of the $n$th row of $\mathbf{X}$ and $\mathbf{a}_n$ represents the transpose of the $n$th row of $\mathbf{A}$. Details of those posterior parameters can be found in the APPENDIX. As the conditional distributions of $\mathbf{A}$ and $\mathbf{S}$ are represented as truncated Gaussian forms, a straightforward Gibbs sampler can be conveniently employed. Sampling from a truncated Gaussian is achieved by employing the method of (Damien and Walker, 2001). We now have a means of sampling from the posterior $p(\mathbf{S}, \mathbf{A}, \mathbf{\Lambda}|\mathbf{X}, \gamma, M)$ the following section now considers how the marginal likelihood can be obtained.

## 3 MODEL COMPLEXITY

### 3.1 BAYESIAN INFORMATION CRITERION

The BIC of (Schwarz, 1978) has been widely used in model selection problems. The BIC, which is an asymptotic expansion of the quantity $\log \int p(\mathbf{X}, \mathbf{\Theta}|M) d\mathbf{\Theta}$, is computed as follows,

$$BIC = \log p(\mathbf{X}|\hat{\mathbf{\Theta}}, M) - \frac{1}{2} k_M \log N$$

where $\hat{\mathbf{\Theta}} = \text{argmax}_{\mathbf{\Theta}} \{\log p(\mathbf{X}|\mathbf{\Theta}, M)\}$, and $k_M = N \times M + M \times T + N$ is the number of parameters to be estimated. In this paper we substitute the $\hat{\mathbf{\Theta}}$ by the output of an NMF algorithm which provides a solution of maximizing the joint likelihood. The model with the largest value of BIC is preferred, and one feature of BIC is that it is independent of model parameter priors.

### 3.2 THERMODYNAMIC INTEGRATION

(Friel and Pettitt, 2008) proposed to estimate the marginal likelihood via power posteriors which is based on ideas of thermodynamic integration (TI) or path sampling (Gelman and Meng, 1998). Compared to

BIC, thermodynamic integration is sensitive to model parameter priors, which could be useful for selecting parameter priors for models. Consider a temperature parameter $t \in [0,1]$, based on the power posterior $p_t(\mathbf{\Theta}|\mathbf{X},M) \propto p(\mathbf{X}|\mathbf{\Theta},M)^t p(\mathbf{\Theta})$, the thermodynamic integral follows in the form (Friel and Pettitt, 2008),

$$\log p(\mathbf{X}|M) = \int_0^1 \mathbf{E}_{p_t}\{\log p(\mathbf{X}|\mathbf{\Theta},M)\}dt$$

where $\mathbf{E}_{p_t}$ denotes the mathematical expectation with respect to $p_t(\mathbf{\Theta}|\mathbf{X},M)$. Choosing an appropriate discretization $0 = t_0 < t_1 < \cdots < t_{n-1} < t_n = 1$, this integral can be estimated by using e.g. trapezoidal integration to obtain a numerical approximation of the above integral then $\log p(\mathbf{X}|M) \approx \frac{1}{2}\sum_{i=0}^{n-1}(t_{i+1} - t_i)\mathbf{E}_{p_{t_{i+1}}}\{\log p(\mathbf{X}|\mathbf{\Theta},M)\} + \mathbf{E}_{p_{t_i}}\{\log p(\mathbf{X}|\mathbf{\Theta},M)\}$ with the expectations being obtained via Monte Carlo estimates with the form $\mathbf{E}_{p_{t_i}}\{\log p(\mathbf{X}|\mathbf{\Theta},M)\} \approx \frac{1}{K}\sum_{k=1}^{K}\log p(\mathbf{X}|\mathbf{\Theta}^k,M)$ where $\mathbf{\Theta}^k$ denotes a sample drawn from $p_{t_i}(\mathbf{\Theta}|\mathbf{X},M)$. Monte Carlo standard errors for this approximation follow straightforwardly (Friel and Pettitt, 2008). For this numerical approximation we can also compute a corresponding lower and upper bound on the log-marginal likelihood (Calderhead and Girolami, 2008): $\log p(\mathbf{X}|M) \geq \sum_{i=0}^{n-1}(t_{i+1} - t_i)\mathbf{E}_{p_{t_i}}\{\log p(\mathbf{X}|\mathbf{\Theta},M)\}$, $\log p(\mathbf{X}|M) \leq \sum_{i=1}^{n}(t_i - t_{i-1})\mathbf{E}_{p_{t_i}}\{\log p(\mathbf{X}|\mathbf{\Theta},M)\}$. This estimator of the log-marginal likelihood exploits the Gibbs sampler presented previously at each step of the temperature ladder and the marginal likelihoods for each set of candidate models $M$ is then computed to assess the plausible values of $M$. This is found to be quite costly computationally so an RJMCMC approach is now developed.

## 4 RJMCMC FOR NMF

In this section we derive an RJMCMC algorithm which naturally considers $M$ as a discrete random variable. In model (1), for a variable NoC $m$, we denote the parameter space as $\mathbf{\Theta}_m = \{\mathbf{A},\mathbf{S},\mathbf{\Lambda}\}$ where $\mathbf{A}$ and $\mathbf{S}$ are defined as previously. In the setting of RJMCMC, the model indicator $m$ is also considered as an unknown parameter, and thus the whole parameter space can then be written as $\mathcal{C} = \cup_{m=1}^{m_{max}}\{\{m\} \times \mathbf{\Theta}_m\}$ where $m_{max}$ denotes the maximum NoC. In the following context, we denote $\mathcal{C}_m = \{m\} \times \mathbf{\Theta}_m$, and $\boldsymbol{\theta}_m$ as an element of $\mathbf{\Theta}_m$. In the setting of Green's RJMCMC (Green, 1995), the target invariant distribution is the posterior $p(m,\boldsymbol{\theta}_m|\mathbf{X},\gamma)$. When the current state is $\{m,\boldsymbol{\theta}_m\} \in \mathcal{C}_m$, the acceptance probability of moving to state $\{m',\boldsymbol{\theta}_{m'}\} \in \mathcal{C}_{m'}$ is $\alpha(m,m') = \min\left\{1, \frac{p(m',\boldsymbol{\theta}_{m'}|\mathbf{X},\gamma)\pi(m',m)q_m(m,\boldsymbol{\theta}_m)}{p(m,\boldsymbol{\theta}_m|\mathbf{X},\gamma)\pi(m,m')q_{m'}(m',\boldsymbol{\theta}_{m'})}\left|\frac{\partial g(m,\boldsymbol{\theta}_m,m',\boldsymbol{\theta}_{m'})}{\partial(m,\boldsymbol{\theta}_m,m',\boldsymbol{\theta}_{m'})}\right|\right\}$ where $q_{m'}(m',\boldsymbol{\theta}_{m'})$ is a proposal distribution for $(m',\boldsymbol{\theta}_{m'})$, $\pi(m,m')$ is the probability of moving from subspace $\mathcal{C}_m$ to $\mathcal{C}_{m'}$, $g(\cdot)$ denotes a bijective function and in this paper we set it to be an identity function, i.e. $g(y) = y$, and $|\cdot|$ denotes the Jacobian. Clearly in this case the Jacobian is one. Note that for moving from one subspace to another the proposals must satisfy the dimension-matching requirement which has clearly been achieved by the bijective function.

According to (Green, 1995), the proposal $q_m(\cdot)$ could be any distribution as long as it satisfies the dimension-matching requirement. However, in practice a proposal distribution is required so the algorithm jumps from one subspace to another. Devising a good proposal distribution for specific problems is a well known difficulty of RJMCMC (Brooks *et al.*, 2003; Godsill, 2001; Han and Carlin, 2001). For the current problem, it would be possible to add or remove some rows and columns of $\mathbf{S}$ and $\mathbf{A}$ 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 case of $\pi(m',m) = \pi(m,m')$, a good proposal should frequently have acceptance probabilities of values approximately equal to one. Therefore it is able to make the RJMCMC jump back from the next subspace to the current one. So if we set $q_m(m,\theta_m) = p(m,\theta_m|\mathbf{X},\gamma)$, i.e., the proposal distribution is exactly the joint posterior, the acceptance probability is exactly one. However this is not the case as the posterior usually does not have a simple form, and approximation methods should then be employed. One possible solution would employ the Laplace approximation to estimate the modes and variances of the posteriors. In this case, the posterior is approximately represented as a Gaussian form, and thus proposal samples can then be drawn from this Gaussian distribution. For the current problem, all posteriors of those parameters have truncated Gaussian or Gamma distributions so we follow the schemes which have been successfully employed in some other problems for searching for a suitable proposal (Han and Carlin, 2001; Lopes and West, 2004). The bijective function was set to be an identity function, and this means we are employing an independent sampler which works best when the proposal is a reasonable approximation to the target posterior distribution. So before implementing RJMCMC, as in (Han and Carlin, 2001; Lopes and West, 2004) we obtain the posterior of each parameter in each subspace, which are then employed as the proposal distributions. In this paper we only consider the jumps between models $m$ and $m+1$, and the jumps between any two of those models could be straightforward.

Suppose we have obtained $\mathcal{S}$ conditional posterior samples of $\mu_{s_{mt}}, \sigma^2_{s_m}, \mu_{a_{nm}}, \sigma^2_{a_{nm}}, \alpha_n$ and $\beta_n$ as $\xi$. We can obtain the Monte Carlo estimates of those parameters in all the candidate subspaces such that $\hat{\xi} = \frac{1}{\mathcal{S}} \sum_{s=1}^{\mathcal{S}} \xi^{(s)}$, then the proposal of each variable $a_{nm}, s_{mt}$ and $\lambda_n^{-1}$ can be represented as,

$$\begin{aligned} q_m(a_{nm}) &= \mathcal{N}_{a_{nm}}(\hat{\mu}_{a_{nm}}, \hat{\sigma}^2_{a_{nm}}) \mathbf{1}_{[c_a, d_a]}(a_{nm}), \\ q_m(s_{mt}) &= \mathcal{N}_{s_{mt}}(\hat{\mu}_{s_{mt}}, \hat{\sigma}^2_{s_m}) \mathbf{1}_{[c_s, d_s]}(s_{mt}), \\ q_m(\lambda_n^{-1}) &= \texttt{Gamma}(\hat{\alpha}_n, \hat{\beta}_n). \end{aligned}$$

In order to compute the acceptance probability, we need to compute the quantity

$$\begin{aligned} &\log p(m, \theta_m | \mathbf{X}) \pi(m, m') q_{m'}(m', \theta_{m'}) \\ &= \log p(m, \theta_m | \mathbf{X}) + \log \pi(m, m') + \log q_{m'}(m', \theta_{m'}) \\ &\propto \log p(\mathbf{X} | m, \theta_m, \gamma) + \log p(\mathbf{A}, \mathbf{S}, \Lambda | m, \gamma) + \log p(m) \\ &\quad + \log \pi(m, m') + \sum\nolimits_{m',t} \log q_{m'}(s_{m't}) \\ &\quad + \sum_{n,m'} \log q_{m'}(a_{nm'}) + \sum_n \log q_{m'}(\lambda_n^{-1}) \end{aligned}$$

So each of the above items are represented as:

$$\begin{aligned} \log p(\mathbf{X} | m, \theta_m, \gamma) &= -\frac{1}{2} TN \log 2\pi - \frac{1}{2} T \sum_{n=1}^{N} \log \lambda_n \\ &\quad - \frac{1}{2} \sum_{t=1}^{T} (\mathbf{x}_t - \mathbf{A}\mathbf{s}_t)^T \mathbf{\Lambda}^{-1} (\mathbf{x}_t - \mathbf{A}\mathbf{s}_t) \\ \log p(\mathbf{A}, \mathbf{S}, \mathbf{\Lambda} | m, \gamma) &= MT \log (d_s - c_s)^{-1} + NM \log \alpha_a \\ &\quad - \alpha_a \sum_{n,m} a_{nm} + (\alpha - 1) \sum_{n=1}^{N} \log \lambda_n^{-1} - \frac{1}{\beta} \sum_{n=1}^{N} \lambda_n^{-1} \\ \log q_{m'}(a_{nm'}) &= -\frac{1}{2} \left\{ \log 2\pi \hat{\sigma}^2_{a_{nm'}} + \frac{(a_{nm'} - \hat{\mu}_{a_{nm'}})^2}{\hat{\sigma}^2_{a_{nm'}}} \right\} \\ \log q_{m'}(s_{m't}) &= -\frac{1}{2} \left\{ \log 2\pi \hat{\sigma}^2_{s_{m'}} + \frac{(s_{m't} - \hat{\mu}_{s_{m't}})^2}{\hat{\sigma}^2_{s_{m'}}} \right\} \\ \log q_{m'}(\lambda_n^{-1}) &= (\hat{\alpha}_n - 1) \log \lambda_n^{-1} - (\hat{\beta}_n \lambda_n)^{-1} \\ &\quad - \hat{\alpha}_n \log \hat{\beta}_n - \log \Gamma(\hat{\alpha}_n) \end{aligned}$$

Note that in this paper we always suppose the prior probability of moving from one subspace to another is 0.5, i.e., $\pi(m, m') = 0.5$, and the prior for $m$ is a discrete uniform distribution with support $\{m_{min}, m_{min}+1, \ldots, m_{max}\}$ where $m_{min}$ denotes the minimum NoC. Suppose the current state is in the subspace $\mathcal{C}_m$, and we only consider jumps to $\mathcal{C}_{m-1}$ which is denoted by a DEATH step and $\mathcal{C}_{m+1}$ which is denoted by a BIRTH step. The probability of attempting a BIRTH or DEATH step when the current state is in $\mathcal{C}_m$ is given by $b_m$ and $d_m$, respectively. In this paper, we set $b_{m_{max}} = d_{m_{min}} = 0$, $d_{m_{max}} = b_{m_{min}} = 1$ and $b_m = d_m = 0.5$ for all other values of $m$. The RJMCMC algorithm can then be represented as follows:

- 1. Obtain proposal distributions
- 2. Initialize $\mathbf{A}, \mathbf{S}$, and $\mathbf{\Lambda}$, and set the current model indicator as $m$. Set $\pi(m, m') = 0.5$.
- 3. Main loop of RJMCMC algorithm.
  - 1. Sample $\mathbf{A}, \mathbf{S}$, and $\mathbf{\Lambda}$.
  - 2. If $m = m_{min}$, then perform BIRTH step.
  - 3. If $m = m_{max}$, then perform DEATH step.
  - 4. If $m_{min} < m < m_{max}$, draw a uniform random variable $u \sim U(0, 1)$.
    - 1. If $u \leq b_m$, then perform BIRTH step;
    - 2. else if $u \leq b_m + d_m$, then perform DEATH step;
  - 5. Repeat.

The BIRTH step is described as follows, and the DEATH step is similar:

- 1. Draw samples from proposal distributions $q_{m+1}$ in $\mathcal{C}_{m+1}$ subspace.
- 2. Compute acceptance probability $\alpha(m, m+1)$.
- 3. Draw a uniform random variable $u \sim U(0, 1)$.
- 4. If $u \leq \alpha(m, m+1)$, then accept the proposal state and set the next state model indicator to be $m + 1$.
- 5. Else set the next state to the current state.

We now have the required methodology to tackle our main problem and firstly use a well known toy problem to illustrate the various approaches presented.

# 5 EXPERIMENTS

## 5.1 SYNTHETIC IMAGE DATA

In this section we apply the described methods to some synthetic image data which were generated by using the following mixing matrices (of dimension 7 × 2, 7 × 3 and 7 × 4) and the component images are shown in the first row of figure 1,

$$\begin{pmatrix} 1.0 & 0.3 \\ 0.3 & 1.0 \\ 1.0 & 1.0 \\ 0.3 & 0.3 \\ 0.1 & 1.0 \\ 1.0 & 0.1 \\ 0.1 & 0.1 \end{pmatrix}, \begin{pmatrix} 1.0 & 1.0 & 1.0 \\ 1.0 & 0.3 & 0.3 \\ 0.3 & 1.0 & 0.3 \\ 0.3 & 0.3 & 1.0 \\ 1.0 & 1.0 & 0.0 \\ 0.0 & 1.0 & 1.0 \\ 1.0 & 0.0 & 1.0 \end{pmatrix}, \begin{pmatrix} 1.0 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1.0 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1.0 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1.0 \\ 1.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 1.0 \end{pmatrix}$$

Gaussian noise with zero mean and variance levels of 0.05, 0.01, 0.001, 0.0001 were

Table 1: Estimated number of components by using BIC, TI and RJMCMC.

| | 2 components | | | |
|---|---|---|---|---|
| Noise Level | 0.05 | 0.01 | 0.001 | 0.0001 |
| BIC | 2 | 2 | 2 | 2 |
| TI | 2 | 2 | 2 | 2 |
| RJMCMC | 2 | 2 | 2 | 2 |
| | 3 components | | | |
| Noise Level | 0.05 | 0.01 | 0.001 | 0.0001 |
| BIC | 4 | 4 | 6 | 6 |
| TI | 3 | 3 | 3 | 3 |
| RJMCMC | 3 | 3 | 3 | 3 |
| | 4 components | | | |
| Noise Level | 0.05 | 0.01 | 0.001 | 0.0001 |
| BIC | 5 | 7 | 7 | 7 |
| TI | 4,5 | 4,5 | 4,5 | 4,5 |
| RJMCMC | 2,3,4 | 4 | 4 | 4 |

Table 2: Computing time in seconds for number of iterations (NoI) after which the algorithms converged. Standard deviation values are shown in brackets.

| | Image Data | | |
|---|---|---|---|
| Method | BIC | TI | RJMCMC |
| NoI | 50000 | 1000 | 20000 |
| Time | 826(36) | 4.5e+3(95) | 5.3e+3(368) |
| | Experimental Data | | |
| Method | BIC | TI | RJMCMC |
| NoI | 100000 | 20000 | 250000 |
| Time | 7.3e+3 | 5.7e+5 | 6.4e+4 |

added to the synthetic image combinations. The images used here were downloaded from http://www.cs.helsinki.fi/u/phoyer/NCimages.html. Before working on model selection, we need to diagnose the convergence of the Gibbs sampler, which was achieved by using the potential scale reduction factor (PSRF) of (Gelman and Rubin, 1992) which is usually denoted by $\hat{R}$. Here we use the iterated graphical approach described in (Brooks and Gelman, 1998) to compute $\hat{R}$. The method divides the $I$ sequences with length $2T$ into batches of length $b$, which gives a series of sequences of lengths $2kb$ where $k = 1, 2, \cdots, T/b$. We then compute $\hat{R}(k)$ using the latter half of the $k$th sequence. We compute $\hat{R}$ for the seven noise variances with $b = 100$ and suppose the Gibbs sampler was converged after all the $\hat{R}$ were less than 1.2. The results show that the Gibbs sampler converged after 500 iterations. We then applied the BIC, TI and RJMCMC to these image data to infer the number of component images. We used the maximized joint likelihood outputs of NMF to compute BIC. For TI, the temperature parameters were set to $t_i = (i/20)^3$ where $i = 0, 1, \cdots, 20$, and the Gibbs samplers ran 20000 iterations with the last 10000 being used to compute the marginal likelihood. For RJMCMC, we also need to diagnose the convergence and here we employ the method of (Brooks and Giudici, 2000) which monitors some particular function of the parameters such as the log likelihood. We run $I$ chains with $2T$ iterations and then compute the total variations both between chains and between models. Some quantities are then computed and they are the total variation $\hat{V}$, the within-chain variance $W_c$, the within-model variance $W_m$, the variance within both chains and models $W_m W_c$, the between-model variance $B_m$, and the within-chain variation split between and averaged over models $B_m W_c$. Details of these can be found in (Brooks and Giudici, 2000). Essentially the ratio $\hat{V}/W_c$ is the $\hat{R}$ of Gelman and Rubin. Comparing $W_m$ and $W_m W_c$, which should well approximate the true mean within-model variance, tells us how well the chains are mixing within models, and whilst comparing $B_m$ and $B_m W_c$, which should well approximate the true between-model variance, tells us how well the chains are mixing between models. Five chains each running 100000 iterations were used to diagnose the convergence of the RJMCMC and we employ the log-likelihood as the scalar parameter for diagnosing convergence which has also been used by (Brooks and Giudici, 2000). The results showed that after 20000 iterations all these quantities were mixing very well. We actually used the second half of the chains of length 100000 iterations to compute the model posterior distributions. All the results are shown in Table 1. In the case of 7 mixtures with 2 components, both BIC and RJMCMC can correctly locate NoC. For TI, the estimated log-marginal likelihood strongly prefers 2 components. However, if we take into account the Monte Carlo sampling variability of the estimate of the log-marginal likelihood there is no significant evidence that models with $2 \sim 7$ components are different. So we show the most likely NoC estimated by TI in Table 1. In the case of 7 mixtures with 3 components, BIC failed to locate NoC whilst both TI and RJMCMC can correctly estimate NoC. In the case of 7 mixtures with 4 components, BIC again failed to locate NoC. TI prefers 4 and 5 components. RJMCMC suggested $2 \sim 4$ components when the noise level was 0.05 and in the other cases it correctly estimates NoC. After the NoC was located, the components can then be recovered using the Gibbs sampler. Figure 1 shows an example. We also compare these methods providing computing times, and the results are shown in Table 2. Note that the experiments were done using Matlab-7.4 under Centos Linux 5 (64-bit, kernel 2.6.18), running on a

2×dual-core AMD 2214 CPU which runs at 2.2 GHz, and the machine has 4GB memory. In the table, the number of iterations show the burn-in required for convergence. We collected the time of each case shown in Table 1, and then the mean and standard deviation values are shown in Table 2. It shows that BIC requires less computing time as a simple gradient based method is used in optimization, and as expected TI and RJMCMC require more time as they are both sampling based methods.

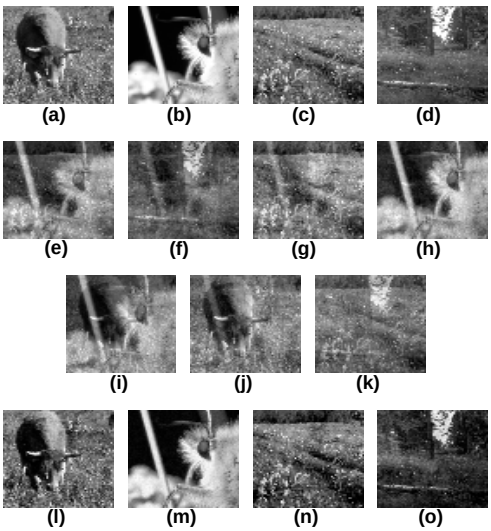


Figure 1: (a-d): Original images; (e-k): Mixtures of images; (l-o): Posterior mean pixel values (inferred images) from the Gibbs sampler.

### 5.2 MULTIPLEXED RAMAN SPECTRA

In this section we applied the described methods to Surface-Enhanced Resonance Raman Scattering (SERRS) data (Graham *et al.*, 2006). This novel data set was generated from multiplexed detection of six dye labelled oligonucleotides using SERRS (Faulds *et al.*, 2004). The purpose of this experiment was the simultaneous detection of six different DNA sequences corresponding to different strains of the Escherichia coli bacterium, each labelled with a different commercially available dye label. The six dye labels were TAMRA, ROX, HEX, TET, FAM and Cy3 (see their SERRS spectra in the first row of figure 4). The multiplexing was carried out using every possible combination of the six labeled oligonucleotides resulting in 64 sample spectra. Each sample was represented by an SERRS spectrum containing 574 points. We assume these measured Raman spectra are a linear mixture of the spectra of the independent dye labels. So if

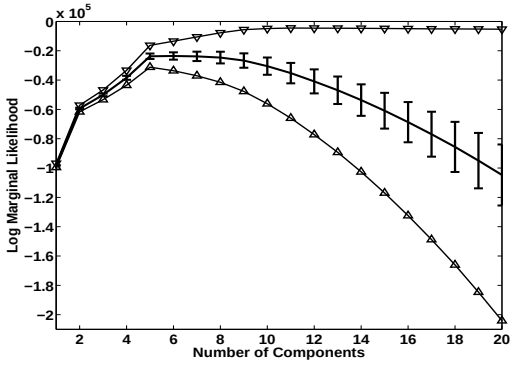


Figure 2: A numerical approximation and an overall Monte Carlo standard error (middle line), upper ($\bigtriangledown$) and lower ($\bigtriangleup$) bounds of log marginal likelihood of each model computed by using the thermodynamic integration.

we represent the Raman spectra as $\mathbf{X} \in \mathcal{R}_{+}^{64\times 574}$, we then decompose it into several individual spectra as is shown in equation (1) where $N = 64$ and $T = 574$. In order to recover those component spectra, we firstly need to estimate the NoC. BIC results using the maximized likelihood outputs of NMF suggested the NoC was around 10 which is far from the correct NoC of six. Before performing TI and RJMCMC, we also need to diagnose the convergence of the Gibbs sampler. We employed the 64 noise variances to diagnose convergence and it showed that after 20000 iterations the Gibbs sampler had converged. For the TI, the temperature parameters were set to $t_i = (i/20)^3$ where $i = 0, 1, \cdots, 20$, and the Gibbs samplers ran 40000 iterations with the last 20000 being used to compute the marginal likelihood. The final log-marginal likelihoods of each model are shown in figure 2. This would suggest the NoC lie in the plausible range of $5 \sim 7$. For the RJMCMC, we again employ the log-likelihood as the scalar parameter for diagnosing convergence. Five chains with each running 1000000 iterations started from randomly generated points were used to diagnose the convergence. We set $b = 25000$. The results are shown in figure 3. It shows that both $\hat{V}$ and $W_c$ approximated the true variance well, and the plots of $W_m$ and $W_m W_c$ show that the chains were mixing very well within models after 250000 samples. The plots of $B_m$ and $B_m W_c$ show that after 250000 iterations, the chains were well mixed between models which is essential for computing the posterior model probabilities. The posterior distributions of the NoC computed using the second half of these five chains are plotted in figure 3 (d). This suggests the NoC should lie on the plausible range of $5 \sim 7$. Combining all these results we concluded that the posterior for $m$ mostly favors $5 \sim 7$ component spectra, which seems reasonable and

covers the true value of six components. In the case of $m = 5$, figure 4 (g-k) shows the recovered spectra, and it shows that ROX and TAMRA are likely to merge into one spectrum (j) since both spectra have peaks in similar frequencies (see plot (l)). We indicate that the only way to differentiate spectra is to look at frequencies (horizontal axis of the plots in figure 4) where peaks appear. In the case of $m = 7$, figure 4 (s-y) plots the recovered spectra, and it is likely that there is one noise component additionally. Figure 4 (m-r) shows the recovered spectra when $m = 6$, which are comparable to the original ones. The time for each method to converge is shown again in Table 2. Now for this larger problem, it shows that BIC is again the fastest method, but it failed to locate reasonable NoC, and TI became slower than RJMCMC which is due to the increase in the number of possible components.

These results are highly encouraging as the RJMCMC and TI methods have been demonstrated to be capable of infering the plausible numbers of components as well as inferring the component spectra from the multiplexed readouts.

## 6 CONCLUSIONS

A fully Bayesian framework for the analysis of NMF has been proposed. A crucial task which has largely been overlooked in the literature is the estimation of the underlying number of components. We have considered several methods namely BIC and TI, and developed RJMCMC for the NMF problem. The results have shown that BIC has failed to locate NoC for the data employed. TI and RJMCMC were able to locate the correct NoC, which are suitable for model selection for NMF. For the experimental Raman spectra, when the NoC was located we have shown that the component spectra were able to be inferred from the corresponding posteriors. This indicates that the NMF is suitable for analysing Raman spectra. It has not escaped our attention that the TI and RJMCMC methods we have proposed could also be used for estimating the number of sources for standard independent component analysis e.g.(Fevotte and Godsill, 2006).

## A APPENDIX - CONDITIONAL DISTRIBUTIONS

In this section we give the derivations of the full posterior distribution $p(\mathbf{S}, \mathbf{A}, \mathbf{\Lambda}|\mathbf{X}, \gamma)$, which is proportional to

$$\begin{aligned} & p(\mathbf{X}|\mathbf{S}, \mathbf{A}, \mathbf{\Lambda}) p(\mathbf{A}|\alpha_a) p(\mathbf{S}|c_s, d_s) p(\mathbf{\Lambda}|\alpha_\lambda, \beta_\lambda) \\ = & \prod_t \mathcal{N}_{\mathbf{x}_t}(\mathbf{A}\mathbf{s}_t, \mathbf{\Lambda}) \prod_{n,m} p(a_{nm}) \prod_{m,t} p(s_{mt}) \prod_n p(\lambda_n^{-1}) \end{aligned}$$

The conditional distribution of each $s_{mt}$ can then be easily obtained with the following form,

$$\begin{aligned} & p(s_{mt}|\mathbf{x}_t, \mathbf{A}, \mathbf{s}_{-m,t}, \mathbf{\Lambda}) \propto p(\mathbf{x}_t|\mathbf{s}_t, \mathbf{A}, \mathbf{\Lambda}) p(s_{mt}|c_s, d_s) \\ \propto & \exp\left\{-\frac{1}{2}\left(A_{s_m} s_{mt}^2 - 2B_{s_{mt}} s_{mt}\right)\right\} \mathbf{1}_{[c_s,d_s]}(s_{mt}) \\ \propto & \mathcal{N}_{s_{mt}}(\mu_{s_{mt}}, \sigma_{s_m}^2) \mathbf{1}_{[c_s,d_s]}(s_{mt}) \end{aligned}$$

where $\mathbf{s}_{-m,t}$ represents all the elements of vector $\mathbf{s}_t$ excluding $s_{mt}$, $\mu_{s_{mt}} = A_{s_m}^{-1} B_{s_{mt}}$ and $\sigma_{s_m}^2 = A_{s_m}^{-1}$ where $A_{s_m} = \sum_{i=1}^N \frac{a_{im}^2}{\lambda_i}$ and $B_{s_{mt}} = \sum_{i=1}^N \frac{a_{im}}{\lambda_i}\left(x_{it} - \sum_{j=1, j\neq m}^M a_{ij} s_{jt}\right)$. Similarly, the conditional distribution of $a_{nm}$ is represented as,

$$\begin{aligned} & p(a_{nm}|\mathbf{x}_n, \mathbf{S}, \mathbf{a}_{n,-m}, \lambda_n) \propto p(\mathbf{x}_n|\mathbf{a}_n, \mathbf{S}, \lambda_n) p(\mathbf{a}_n|\alpha_a) \\ \propto & \exp\left\{-\frac{1}{2}(A_{a_{nm}} a_{nm}^2 - 2B_{a_{nm}} a_{nm})\right\} \exp\{-\alpha_a a_{nm}\} \\ \propto & \mathcal{N}_{a_{nm}}(\mu_{a_{nm}}, \sigma_{a_{nm}}^2) \mathbf{1}_{[c_a,d_a]}(a_{nm}) \end{aligned}$$

where $\mathbf{a}_{n,-m}$ represents all the elements of vector $\mathbf{a}_n$ excluding $a_{nm}$, $\mu_{a_{nm}} = A_{a_{nm}}^{-1}(B_{a_{nm}} - \alpha_a)$ and $\sigma_{a_{nm}}^2 = A_{a_{nm}}^{-1}$ where $A_{a_{nm}} = \lambda_n^{-1}\left(\sum_{t=1}^T s_{mt}^2\right)$ and $B_{a_{nm}} = \lambda_n^{-1}\left(\sum_{t=1}^T s_{mt}\left(x_{nt} - \sum_{j=1, j\neq m}^M a_{nj} s_{jt}\right)\right)$. Finally,

$$\begin{aligned} p(\lambda_n^{-1}|\mathbf{x}_n, \mathbf{S}, \mathbf{a}_n) & \propto p(\mathbf{x}_n|\mathbf{a}_n, \mathbf{S}) p(\lambda_n^{-1}|\alpha_\lambda, \beta_\lambda) \\ & \propto \texttt{Gamma}(\alpha_n, \beta_n) \end{aligned}$$

where $\beta_n = \left(\beta_\lambda + \frac{1}{2}\sum_{t=1}^T \left(x_{nt} - \mathbf{a}_n^T \mathbf{s}_t\right)^2\right)^{-1}$ and $\alpha_n = \alpha_\lambda + \frac{T}{2}$. These conditional distributions give those of the vector forms shown in section 2.

**Acknowledgements**

M. Zhong & M. Girolami are supported by Engineering and Physical Sciences Research Council (EPSRC) basic technology grant EP/E032745/1 "The Molecular Nose". M. Girolami is funded by an EPSRC Advanced Research Fellowship EP/E052029/1.

## References

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring conergence of iterative simulations. *J. Comp. Graph. Stat.*, **7**(4), 434–455.

Brooks, S. P. and Giudici, P. (2000). Markov chain Monte Carlo convergence assessment via two-way analysis of variance. *J. Comp. Graph. Stat.*, **9**(2), 266–285.

Brooks, S. P., Giudici, P., and Roberts, G. O. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *J. R. Statist. Soc. B*, **65**(1), 3–39.

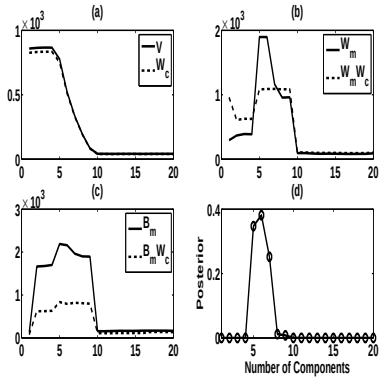


Figure 3: Diagnostic plots for the experimental data. (a) Plot of $\hat{V}$ and $W_c$. (b) Plot of $W_m$ and $W_m W_c$. (c) Plot of $B_m$ and $B_m W_c$. (d) Posterior distribution of the number of component spectra $m$ computed by using the RJMCMC.

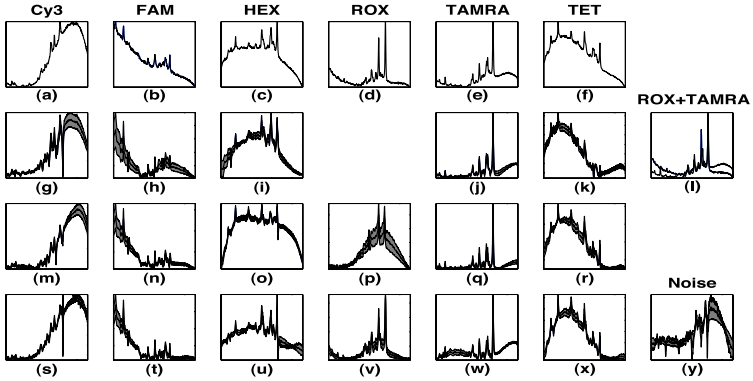


Figure 4: (a-f): The six component spectra; (g-k): The five extracted spectra; (l): Plots of ROX and TAMRA showing that their peaks appear at similar frequencies (horizontal axis), which indicates that ROX and TAMRA would merge into one spectrum (j); (m-r): The six extracted spectra; (s-y): The seven extracted spectra where (y) seems to be noise. The shaded area represents the 95% confidence region of the posterior.


Calderhead, B. and Girolami, M. (2008). Estimating Bayes factors for nonlinear ODE models via thermodynamic integration and population MCMC. Technical report, Department of Computing Science, University of Glasgow.

Damien, P. and Walker, S. G. (2001). Sampling truncated Normal, Beta, and Gamma densities. *J. Comp. Graph. Stat.*, **10**(2), 206–215.

Faulds, K., Smith, W., and Graham, D. (2004). Evaluation of surface-enhanced resonance raman scattering for quantitative DNA analysis. *Anal. Chem.*, **76**(2), 412–417.

Fevotte, C. and Godsill, S. J. (2006). A Bayesian approach for blind separation of sparse sources. *IEEE Trans. Speech Audio Proc.*, **14**(6), 2174–2188.

Friel, N. and Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *J. R. Statist. Soc. B*, **70**(3), 589–607.

Gelman, A. and Meng, X. L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.*, **13**(2), 163–185.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.*, **7**(4), 457–511.

Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *J. Comp. Graph. Stat.*, **10**(2), 230–248.

Graham, D., Faulds, K., and Smith, E. (2006). Biosensing using silver nanoparticles and surface enhanced resonance raman scattering. *Chem. Commun.*, **42**, 4363–4371.

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

Han, C. and Carlin, B. (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *J. Amer. Stat. Assoc.*, **96**(455), 1122–1132.

Hoyer, P. (2004). Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, **5**, 1457–1469.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, volume 13, pages 556–562. MIT Press.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, **14**, 41–67.

Moussaoui, S., Brie, D., Mohammad-Djafari, A., and Carteret, C. (2006). Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling. *IEEE Trans. Sig. Proc.*, **54**(11), 4133–4145.

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

Raftery, A. E., Newton, M. A., Satagopan, J. M., and Krivitsky, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics*, **8**, 1–45.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, **6**(2), 461–464.