Reversible Jump MCMC for Non-Negative Matrix Factorization

Mark Girolami and Mingjun Zhong

Department of Computing Science
University of Glasgow
http://www.dcs.gla.ac.uk/inference/
{mingjun, girolami}@dcs.gla.ac.uk

June 22, 2009
Introduction
Outline

- Introduction
- Non-Negative Matrix Factorization (NMF)
Outline

- Introduction
- Non-Negative Matrix Factorization (NMF)
- Model Selection Problem
Outline

- Introduction
- Non-Negative Matrix Factorization (NMF)
- Model Selection Problem
- Gibbs Sampler for NMF
Outline

- Introduction
- Non-Negative Matrix Factorization (NMF)
- Model Selection Problem
- Gibbs Sampler for NMF
- Model Selection Methods
Outline

- Introduction
- Non-Negative Matrix Factorization (NMF)
- Model Selection Problem
- Gibbs Sampler for NMF
- Model Selection Methods
- Experiments
Outline

- Introduction
- Non-Negative Matrix Factorization (NMF)
- Model Selection Problem
- Gibbs Sampler for NMF
- Model Selection Methods
- Experiments
- Conclusions
Outline

- Introduction
- Non-Negative Matrix Factorization (NMF)
- Model Selection Problem
- Gibbs Sampler for NMF
- Model Selection Methods
- Experiments
- Conclusions
- References
We consider the \textit{model selection problem} for NMF.
We consider the **model selection problem** for NMF.

**Mixture Images**
We consider the **model selection problem** for NMF.
We consider the model selection problem for NMF.

Mixture Images

Non-Negative Matrix Factorization

How many images to be extracted?
Given a data matrix $X \in \mathbb{R}^{N \times T}_{+}$, we consider the NMF problem (Paatero & Tapper, 1994; Lee & Seung, 2001)

$$X = AS + E$$
Non-Negative Matrix Factorization

Given a data matrix $X \in \mathbb{R}^{N \times T}_+$, we consider the NMF problem (Paatero & Tapper, 1994; Lee & Seung, 2001)

$$X = AS + E$$

where

* $A \in \mathbb{R}^{N \times M}_+$, $S \in \mathbb{R}^{M \times T}_+$
Given a data matrix $X \in \mathbb{R}_{+}^{N \times T}$, we consider the NMF problem (Paatero & Tapper, 1994; Lee & Seung, 2001)

$$X = AS + E$$

where

* $A \in \mathbb{R}_{+}^{N \times M}$, $S \in \mathbb{R}_{+}^{M \times T}$
* $E \in \mathbb{R}_{+}^{N \times T}$ is the tolerance within each column
Non-Negative Matrix Factorization

Given a data matrix \( \mathbf{X} \in \mathbb{R}^{N \times T} \), we consider the NMF problem (Paatero & Tapper, 1994; Lee & Seung, 2001)

\[
\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{E}
\]

where

* \( \mathbf{A} \in \mathbb{R}_+^{N \times M}, \mathbf{S} \in \mathbb{R}_+^{M \times T} \)
* \( \mathbf{E} \in \mathbb{R}^{N \times T} \) is the tolerance within each column
* Each column of \( \mathbf{E} \) is assumed to follow a Normal distribution with zero mean and covariance \( \mathbf{\Lambda} = \text{diag}(\lambda_1, \cdots, \lambda_N) \).
Given a data matrix $\mathbf{X} \in \mathbb{R}_{+}^{N \times T}$, we consider the NMF problem (Paatero & Tapper, 1994; Lee & Seung, 2001)

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{E}$$

where

* $\mathbf{A} \in \mathbb{R}_{+}^{N \times M}$, $\mathbf{S} \in \mathbb{R}_{+}^{M \times T}$
* $\mathbf{E} \in \mathbb{R}_{+}^{N \times T}$ is the tolerance within each column
* Each column of $\mathbf{E}$ is assumed to follow a Normal distribution with zero mean and covariance $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_N)$.

So what is the aim of NMF?
Non-Negative Matrix Factorization

- NMF is to recover $A \in \mathbb{R}^{N \times M}_{+}$ and $S \in \mathbb{R}^{M \times T}_{+}$. 

Mark Girolami and Mingjun Zhong

Reversible Jump MCMC for Non-Negative Matrix Factorization
Non-Negative Matrix Factorization

- NMF is to recover $A \in \mathbb{R}_+^{N \times M}$ and $S \in \mathbb{R}_+^{M \times T}$.
- However how to decide the $M$ here?
Non-Negative Matrix Factorization

- NMF is to recover $\mathbf{A} \in \mathbb{R}_+^{N \times M}$ and $\mathbf{S} \in \mathbb{R}_+^{M \times T}$.
- However, how to decide the $M$ here?
- Inferring $M$ is the model selection problem.
We need to compute the posterior of $M$ given data $X$, 

We need to compute the posterior of $M$ given data $X$, 

$$P(M|X) = \frac{p(X|M)P(M)}{\sum_m p(X|m)P(M)}$$
We need to compute the **posterior** of $M$ given data $X$,

$$P(M|X) = \frac{p(X|M)P(M)}{\sum_m p(X|m)P(M)}$$

which is to compute the **marginal likelihood** $p(X|M)$. 
We need to compute the **posterior** of $M$ given data $X$,

$$P(M|X) = \frac{p(X|M)P(M)}{\sum_m p(X|m)P(M)}$$

which is to compute the **marginal likelihood** $p(X|M)$.

In this work, we attempt to compute the **posterior** and **marginal likelihood**.
We need to compute the posterior of $M$ given data $X$,

$$P(M|X) = \frac{p(X|M)P(M)}{\sum_m p(X|m)P(M)}$$

which is to compute the marginal likelihood $p(X|M)$.

In this work, we attempt to compute the posterior and marginal likelihood.

Firstly, we need to look at a Gibbs sampler for a fixed $M$ for NMF.
Likelihood

\[ p(X|A, S, \Lambda) = \prod_{t=1}^{T} N_{x_t}(As_t, \Lambda) \]
**Likelihood**

\[
p(X|A, S, \Lambda) = \prod_{t=1}^{T} N_{x_t}(As_t, \Lambda)
\]

**Prior**

\[
p(s_{mt}|c_s, d_s) = \text{Unif}(c_s, d_s)
\]

\[
p(a_{nm}|\alpha_a, c_a, d_a) = \text{Expon}(\alpha_a) \mathbf{1}_{[c_a, d_a]}(a_{nm})
\]

\[
p(\lambda^{-1}_n|\alpha_{\lambda}, \beta_{\lambda}) = \text{Gamma}(\alpha_{\lambda}, \beta_{\lambda})
\]
We need to derive the full conditional distribution,

\[
p(S, A, \Lambda | X, \gamma) \propto p(X | S, A, \Lambda) p(A | \alpha_a) p(S | c_s, d_s) p(\Lambda | \alpha_\lambda, \beta_\lambda) \\
= \prod_t \mathcal{N}_{x_t}(As_t, \Lambda) \prod_{n,m} p(a_{nm}) \prod_{m,t} p(s_{mt}) \prod_n p(\lambda_n^{-1})
\]
The conditional distribution of $s_{mt}$ is

$$p(s_{mt}|x_t, A, s_{-m,t}, \Lambda) \propto p(x_t|s_t, A, \Lambda)p(s_{mt}|c_s, d_s)$$

$$\propto \exp \left\{ -\frac{1}{2} \left( A_{sm}s_{mt}^2 - 2B_{smt}s_{mt} \right) \right\} 1_{[c_s,d_s]}(s_{mt})$$

$$\propto \mathcal{N}_{s_{mt}}(\mu_{s_{mt}}, \sigma_{s_{mt}}^2) 1_{[c_s,d_s]}(s_{mt})$$

where $s_{-m,t}$ represents all the elements of vector $s_t$ excluding $s_{mt}$, 

$$\mu_{smt} = A_{sm}^{-1} B_{smt} \text{ and } \sigma_{smt}^2 = A_{sm}^{-1} \text{ where } A_{sm} = \sum_{i=1}^{N} \frac{a_{im}^2}{\lambda_i}$$

$$B_{smt} = \sum_{i=1}^{N} \frac{a_{im}}{\lambda_i} \left( x_{it} - \sum_{j=1, j \neq m}^{M} a_{ij} s_{jt} \right).$$
Gibbs Sampler

The conditional distribution of $a_{nm}$ is

\[
p(a_{nm} | x_n, S, a_{n,-m}, \lambda_n)
\propto p(x_n | a_n, S, \lambda_n)p(a_n | \alpha_a)
\propto \exp \left\{ -\frac{1}{2} (A_{anm} a_{nm}^2 - 2B_{anm} a_{nm}) \right\} \exp\{-\alpha_a a_{nm}\}
\propto \mathcal{N}_{anm}(\mu_{anm}, \sigma_{anm}^2)1_{[c_a, d_a]}(a_{nm})
\]

where $a_{n,-m}$ represents all the elements of vector $a_n$ excluding $a_{nm}$, $\mu_{anm} = A_{anm}^{-1} (B_{anm} - \alpha_a)$ and $\sigma_{anm}^2 = A_{anm}^{-1}$ where

\[
A_{anm} = \lambda_n^{-1} \left( \sum_{t=1}^{T} s_{mt}^2 \right) \quad \text{and}
\]

\[
B_{anm} = \lambda_n^{-1} \left( \sum_{t=1}^{T} s_{mt} \left( x_{nt} - \sum_{j=1, j\neq m}^{M} a_{nj} s_{jt} \right) \right).
\]
The conditional distribution of $\lambda_n^{-1}$ is

\[ p(\lambda_n^{-1}|x_n, S, a_n) \propto p(x_n|a_n, S)p(\lambda_n^{-1}|\alpha, \beta) \propto \text{Gamma}(\alpha, \beta_n) \]

where $\beta_n = \left( \beta + \frac{1}{2} \sum_{t=1}^{T} (x_{nt} - a_n^T s_t)^2 \right)^{-1}$ and $\alpha_n = \alpha + \frac{T}{2}$. 
Bayesian Information Criterion (BIC) (Schwarz, 1978)
Model Selection Methods

- Bayesian Information Criterion (BIC) (Schwarz, 1978)
- Thermodynamic Integration (TI) (Friel & Pettitt, 2008)
Model Selection Methods

- Bayesian Information Criterion (BIC) (Schwarz, 1978)
- Thermodynamic Integration (TI) (Friel & Pettitt, 2008)
- Reversible Jump MCMC (RJMCMC) (Green, 1995)
The BIC is to compute

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

where
Bayesian Information Criterion

The BIC is to compute

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

where

* \( \hat{\Theta} = \arg\max_{\Theta} \{ \log p(X|\Theta, M) \} \)
Bayesian Information Criterion

The BIC is to compute

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

where

* \( \hat{\Theta} = \arg\max_{\Theta} \{ \log p(X|\Theta, M) \} \)
* \( k_M = N \times M + M \times T + N \) is the number of parameters to be estimated.
Define a power posterior $p_t(\Theta|X, M) = \frac{1}{z(X|t)} p(X|\Theta, M)^t p(\Theta)$
Define a power posterior $p_t(\Theta | X, M) = \frac{1}{z(X|t)} p(X|\Theta, M)^t p(\Theta)$ where

* $t \in [0, 1]$ is a temperature parameter
Define a power posterior $p_t(\Theta|X, M) = \frac{1}{z(X|t)} p(X|\Theta, M)^t p(\Theta)$ where

* $t \in [0, 1]$ is a temperature parameter

* $z(X|t) = \int_\Theta p(X|\Theta, M)^t p(\Theta)d\Theta$
Define a **power posterior** \( p_t(\Theta|X, M) = \frac{1}{z(X|t)} p(X|\Theta, M)^t p(\Theta) \) where

* \( t \in [0, 1] \) is a temperature parameter

* \( z(X|t) = \int_\Theta p(X|\Theta, M)^t p(\Theta) d\Theta \)

* Then \( z(X|t = 0) = 1 \) and \( z(X|t = 1) = p(X|M) \).
Define a power posterior $p_t(\Theta|X, M) = \frac{1}{z(X|t)} p(X|\Theta, M)^t p(\Theta)$ where

* $t \in [0, 1]$ is a temperature parameter
* $z(X|t) = \int_{\Theta} p(X|\Theta, M)^t p(\Theta) d\Theta$
* Then $z(X|t = 0) = 1$ and $z(X|t = 1) = p(X|M)$.
* So we have

$$\log p(X|M) = \int_0^1 \frac{d}{dt} \log z(X|t)$$
\[
\frac{d}{dt} \log z(X|t) = \frac{d}{dt} \log \int_{\Theta} p(X|\Theta, M)^t p(\Theta) d\Theta
\]
\[ \frac{d}{dt} \log z(X|t) = \frac{d}{dt} \log \int_{\Theta} p(X|\Theta, M)^t p(\Theta) d\Theta \]

\[ = \frac{1}{z(X|t)} \int_{\Theta} \frac{dp(X|\Theta, M)^t}{dt} p(\Theta) d\Theta \]
$$\frac{d}{dt} \log z(\mathbf{X}|t) = \frac{d}{dt} \log \int_{\Theta} p(\mathbf{X}|\Theta, M)^t p(\Theta) d\Theta$$

$$= \frac{1}{z(\mathbf{X}|t)} \int_{\Theta} \frac{dp(\mathbf{X}|\Theta, M)^t}{dt} p(\Theta) d\Theta$$

$$= \frac{1}{z(\mathbf{X}|t)} \int_{\Theta} \frac{dp(\mathbf{X}|\Theta, M)^t}{dt} \frac{p(\mathbf{X}|\Theta, M)^t p(\Theta)}{p(\mathbf{X}|\Theta, M)^t} d\Theta$$
\[
\frac{d}{dt} \log z(X|t) = \frac{d}{dt} \log \int_{\Theta} p(X|\Theta, M)^t p(\Theta) d\Theta \\
= \frac{1}{z(X|t)} \int_{\Theta} \frac{dp(X|\Theta, M)^t}{dt} p(\Theta) d\Theta \\
= \frac{1}{z(X|t)} \int_{\Theta} \frac{dp(X|\Theta, M)^t}{dt} \frac{p(X|\Theta, M)^t p(\Theta)}{p(X|\Theta, M)^t} d\Theta \\
= \int_{\Theta} \frac{d \log p(X|\Theta, M)^t}{dt} \frac{p(X|\Theta, M)^t p(\Theta)}{z(X|t)} d\Theta
\]
\[
\frac{d}{dt} \log z(X|t) = \frac{d}{dt} \log \int p(X|\Theta, M)^t p(\Theta) d\Theta \\
= \frac{1}{z(X|t)} \int \left( \frac{d}{dt} p(X|\Theta, M)^t \right) p(\Theta) d\Theta \\
= \frac{1}{z(X|t)} \int \left( \frac{d}{dt} p(X|\Theta, M)^t \right) \frac{p(X|\Theta, M)^t p(\Theta)}{p(X|\Theta, M)^t} d\Theta \\
= \int \log p(X|\Theta, M) \frac{p(X|\Theta, M)^t p(\Theta)}{z(X|t)} d\Theta
\]
Thermodynamic Integration

\[
\frac{d}{dt} \log z(X|t) = \frac{d}{dt} \log \int_{\Theta} p(X|\Theta, M)^t p(\Theta) d\Theta \\
= \frac{1}{z(X|t)} \int_{\Theta} \frac{dp(X|\Theta, M)^t}{dt} p(\Theta) d\Theta \\
= \frac{1}{z(X|t)} \int_{\Theta} \frac{dp(X|\Theta, M)^t}{dt} \frac{p(X|\Theta, M)^t p(\Theta)}{p(X|\Theta, M)^t} d\Theta \\
= \int_{\Theta} \frac{d \log p(X|\Theta, M)^t}{dt} \frac{p(X|\Theta, M)^t p(\Theta)}{z(X|t)} d\Theta \\
= \int_{\Theta} \log p(X|\Theta, M) \frac{p(X|\Theta, M)^t p(\Theta)}{z(X|t)} d\Theta \\
= \mathbb{E}_{p_t} \{\log p(X|\Theta, M)\}
\]

where \( p_t \) denotes the power posterior.
Thermodynamic Integral can be obtained by integrating both sides with respect to $t$

\[ \int_0^1 \mathbb{E}_{p_t} \{ \log p(X|\Theta, M) \} dt = \int_0^1 \frac{d}{dt} \log z(X|t) \]
Thermodynamic Integration can be obtained by integrating both sides with respect to $t$

$$
\int_0^1 \mathbb{E}_{p_t} \{ \log p(\mathbf{X}|\Theta, M) \} dt = \int_0^1 \frac{d}{dt} \log z(\mathbf{X}|t) = \log z(\mathbf{X}|t = 1) - \log z(\mathbf{X}|t = 0)
$$
Thermodynamic Integral can be obtained by integrating both sides with respect to $t$

$$\int_0^1 E_{p_t} \{ \log p(X|\Theta, M) \} dt = \int_0^1 \frac{d}{dt} \log z(X|t)$$

$$= \log z(X|t = 1) - \log z(X|t = 0)$$

$$= \log p(X|M)$$
Choose a discretization $0 = t_0 < t_1 < \cdots < t_{n-1} < t_n = 1$, numerical approximation to the above integral is then

$$
\log p(X|M) \approx \frac{1}{2} \sum_{i=0}^{n-1} (t_{i+1} - t_i) E_{p_{t_{i+1}}} \{ \log p(X|\Theta, M) \} + E_{p_{t_i}} \{ \log p(X|\Theta, M) \}
$$
Choose a discretization $0 = t_0 < t_1 < \cdots < t_{n-1} < t_n = 1$, numerical approximation to the above integral is then

$$\log p(X|M) \approx \frac{1}{2} \sum_{i=0}^{n-1} (t_{i+1} - t_i) E_{p_{t_{i+1}}} \{ \log p(X|\Theta, M) \} + E_{p_{t_i}} \{ \log p(X|\Theta, M) \}$$

The expectation could be obtained by Monte Carlo estimates,

$$E_{p_{t_i}} \{ \log p(X|\Theta, M) \} \approx \frac{1}{K} \sum_{k=1}^{K} \log p(X|\Theta^k, M)$$

where $\Theta^k$ denotes a sample drawn from $p_{t_i}(\Theta|X, M)$.
Set $s_i$ as the MC standard error for each $E_{p_{t_i}}\{\log p(X|\Theta, M)\}$. The overall MC standard error for $\log p(X|M)$ is

$$
\sqrt{\frac{1}{2}(t_2 - t_1)^2 s_1^2 + \sum_{i=2}^{n-1} \frac{1}{2}(t_{i+1} - t_{i-1})^2 s_i^2 + \frac{1}{2}(t_n - t_{n-1})^2 s_n^2}
$$
Thermodynamic Integration

We can also compute the lower and upper bounds of the numerical approximation

\[
\log p(\mathbf{X}|M) \geq \sum_{i=0}^{n-1} (t_{i+1} - t_i) \mathbb{E}_{p_{t_i}(\Theta|\mathbf{X},M)} \{ \log p(\mathbf{X}|\Theta, M) \}
\]

\[
\log p(\mathbf{X}|M) \leq \sum_{i=1}^{n} (t_i - t_{i-1}) \mathbb{E}_{p_{t_i}(\Theta|\mathbf{X},M)} \{ \log p(\mathbf{X}|\Theta, M) \}
\]
RJMCMC consider $M$ as a random variable, here we use $m$ instead.

Notations:

- For a $m$, $\Theta_m = \{A, S, \Lambda\}$ where $A \in \mathbb{R}^{N \times m}_+$, $S \in \mathbb{R}^{m \times T}_+$. 
RJMCMC consider $M$ as a random variable, here we use $m$ instead. Notations:

- For a $m$, $\Theta_m = \{A, S, \Lambda\}$ where $A \in \mathbb{R}^{N \times m}_{+}$, $S \in \mathbb{R}^{m \times T}_{+}$.
- For a $m$, set $\theta_m$ as an element of $\Theta_m$. 
RJMCMC consider $M$ as a random variable, here we use $m$ instead.

Notations:

- For a $m$, $\Theta_m = \{A, S, \Lambda\}$ where $A \in \mathbb{R}^{N \times m}_+$, $S \in \mathbb{R}^{m \times T}_+$.
- For a $m$, set $\theta_m$ as an element of $\Theta_m$
- For a $m$, set $C_m = \{m\} \times \Theta_m$. 
RJMCMC consider $M$ as a random variable, here we use $m$ instead. Notations:

- For a $m$, $\Theta_m = \{A, S, \Lambda\}$ where $A \in \mathbb{R}^{N \times m}, S \in \mathbb{R}^{m \times T}$.
- For a $m$, set $\theta_m$ as an element of $\Theta_m$.
- For a $m$, set $C_m = \{m\} \times \Theta_m$.
- For all $m$, set $C = \bigcup_{m=m_{\min}}^{m_{\max}} C_m = \bigcup_{m=m_{\min}}^{m_{\max}} \{\{m\} \times \Theta_m\}$. 

Mark Girolami and Mingjun Zhong

Reversible Jump MCMC for Non-Negative Matrix Factorization
Consider the whole space $\mathcal{C}$, the target invariant distribution is the posterior $p(m, \theta_m | X, \gamma)$. 
Consider the whole space $\mathcal{C}$, the target invariant distribution is the posterior $p(m, \theta_m | \mathbf{X}, \gamma)$.

When the current state is $\{m, \theta_m\} \in \mathcal{C}_m$, the acceptance probability $\alpha(m, m')$ of moving to state $\{m', \theta_{m'}\} \in \mathcal{C}_{m'}$ is

$$\min \left\{ 1, \frac{p(m', \theta_{m'} | \mathbf{X}, \gamma)p(m', m)q_m(m, \theta_m)}{p(m, \theta_m | \mathbf{X}, \gamma)p(m, m')q_{m'}(m', \theta_{m'})} \left| \frac{\partial g(m, \theta_m, m', \theta_{m'})}{\partial (m, \theta_m, m', \theta_{m'})} \right\} \right\}$$
Consider the whole space $\mathcal{C}$, the \textit{target invariant distribution} is the posterior $p(m, \theta_m | X, \gamma)$.

When the current state is $\{m, \theta_m\} \in \mathcal{C}_m$, the acceptance probability $\alpha(m, m')$ of moving to state $\{m', \theta_{m'}\} \in \mathcal{C}_{m'}$ is

$$\min \left\{ 1, \frac{p(m', \theta_{m'} | X, \gamma) \pi(m', m) q_m(m, \theta_m)}{p(m, \theta_m | X, \gamma) \pi(m, m') q_{m'}(m', \theta_{m'})} \right\} \left\{ \frac{\partial g(m, \theta_m, m', \theta_{m'})}{\partial (m, \theta_m, m', \theta_{m'})} \right\}$$

where

* $q_{m'}(m', \theta_{m'})$ is a proposal distribution for $(m', \theta_{m'})$. 
Consider the whole space $\mathcal{C}$, the **target invariant distribution** is the posterior $p(m, \theta_m|\mathbf{X}, \gamma)$.

When the current state is $\{m, \theta_m\} \in \mathcal{C}_m$, the acceptance probability $\alpha(m, m')$ of moving to state $\{m', \theta_{m'}\} \in \mathcal{C}_{m'}$ is

$$
\min \left\{ 1, \frac{p(m', \theta_{m'}|\mathbf{X}, \gamma)\pi(m', m)q_m(m, \theta_m)}{p(m, \theta_m|\mathbf{X}, \gamma)\pi(m, m')q_{m'}(m', \theta_{m'})} \left| \frac{\partial g(m, \theta_m, m', \theta_{m'})}{\partial (m, \theta_m, m', \theta_{m'})} \right| \right\}
$$

where

* $q_{m'}(m', \theta_{m'})$ is a proposal distribution for $(m', \theta_{m'})$.
* $\pi(m, m')$ is the probability of moving from subspace $\mathcal{C}_m$ to $\mathcal{C}_{m'}$ which was set to 0.5.
Consider the whole space $\mathcal{C}$, the target invariant distribution is the posterior $p(m, \theta_m | X, \gamma)$. When the current state is $\{m, \theta_m\} \in \mathcal{C}_m$, the acceptance probability $\alpha(m, m')$ of moving to state $\{m', \theta_{m'}\} \in \mathcal{C}_{m'}$ is

$$\min \left\{ 1, \frac{p(m', \theta_{m'} | X, \gamma) \pi(m', m) q_m(m, \theta_m)}{p(m, \theta_m | X, \gamma) \pi(m, m') q_{m'}(m', \theta_{m'})} \left| \frac{\partial g(m, \theta_m, m', \theta_{m'})}{\partial (m, \theta_m, m', \theta_{m'})} \right| \right\}$$

where

* $q_{m'}(m', \theta_{m'})$ is a proposal distribution for $(m', \theta_{m'})$.
* $\pi(m, m')$ is the probability of moving from subspace $\mathcal{C}_m$ to $\mathcal{C}_{m'}$ which was set to 0.5.
* $g(\cdot)$ denotes a bijective function which was set to be an identity function, i.e. $g(y) = y$. 
Consider the whole space $\mathcal{C}$, the target invariant distribution is the posterior $p(m, \theta_m | X, \gamma)$.

When the current state is $\{m, \theta_m\} \in \mathcal{C}_m$, the acceptance probability $\alpha(m, m')$ of moving to state $\{m', \theta_{m'}\} \in \mathcal{C}_{m'}$ is

$$\min \left\{ 1, \frac{p(m', \theta_{m'} | X, \gamma)\pi(m', m)q_m(m, \theta_m)}{p(m, \theta_m | X, \gamma)\pi(m, m')q_{m'}(m', \theta_{m'})} \left| \frac{\partial g(m, \theta_m, m', \theta_{m'})}{\partial (m, \theta_m, m', \theta_{m'})} \right| \right\}$$

where

* $q_{m'}(m', \theta_{m'})$ is a proposal distribution for $(m', \theta_{m'})$.
* $\pi(m, m')$ is the probability of moving from subspace $\mathcal{C}_m$ to $\mathcal{C}_{m'}$ which was set to 0.5.
* $g(\cdot)$ denotes a bijective function which was set to be an identity function, i.e. $g(y) = y$.
* $| \cdot |$ denotes the Jacobian which obviously equals one.
Reversible Jump MCMC

How to define the proposal distributions?
Reversible Jump MCMC

How to define the proposal distributions?

We employ the estimated posteriors by using the Gibbs sampler:

\[
q_m(a_{nm}) = \mathcal{N}_{a_{nm}}(\hat{\mu}_{a_{nm}}, \hat{\sigma}_{a_{nm}}^2)
\]
\[
q_m(s_{mt}) = \mathcal{N}_{s_{mt}}(\hat{\mu}_{s_{mt}}, \hat{\sigma}_{s_{mt}}^2)
\]
\[
q_m(\lambda_n^{-1}) = \text{Gamma}(\hat{\alpha}_n, \hat{\beta}_n)
\]

where \(\hat{\xi}\) is the Monte Carlo estimates using the outputs of Gibbs sampler.
Reversible Jump MCMC

It is required to compute the $\alpha(m, m')$:

$$
\log p(m, \theta_m|X) \pi(m, m') q_{m'}(m', \theta_m) \\
= \log p(m, \theta_m|X) + \log \pi(m, m') + \log q_{m'}(m', \theta_m) \\
\propto \log p(X|m, \theta_m, \gamma) + \log p(A, S, \Lambda|m, \gamma) + \log p(m) \\
+ \log \pi(m, m') \\
+ \sum_{m', t} \log q_{m'}(s_{m't}) + \sum_{n, m'} \log q_{m'}(a_{nm'}) + \sum_n \log q_{m'}(\lambda_n^{-1})
$$
\[ \log p(\mathbf{X}|m, \theta_m, \gamma) = -\frac{1}{2} TN \log 2\pi - \frac{1}{2} T \sum_{n=1}^{N} \log \lambda_n \]

\[ -\frac{1}{2} \sum_{t=1}^{T} (\mathbf{x}_t - \mathbf{A}s_t)^T \Lambda^{-1} (\mathbf{x}_t - \mathbf{A}s_t) \]

\[ \log p(\mathbf{A}, \mathbf{S}, \Lambda|m, \gamma) = MT \log (d_s - c_s)^{-1} + NM \log \alpha_a \]

\[ -\alpha_a \sum_{n,m} a_{nm} + (\alpha - 1) \sum_{n=1}^{N} \log \lambda_n^{-1} - \frac{1}{\beta} \sum_{n=1}^{N} \lambda_n^{-1} \]

\[ \log p(m) = \log \frac{1}{m_{\text{max}} - m_{\text{min}}} \]

\[ \log \pi(m, m') = \log 0.5 \]
\[
\log q_m'(a_{nm'}) = -\frac{1}{2} \left\{ \log 2\pi \hat{\sigma}_{a_{nm'}}^2 + \frac{(a_{nm'} - \hat{\mu}_{a_{nm'}})^2}{\hat{\sigma}_{a_{nm'}}^2} \right\}
\]

\[
\log q_m'(s_{m't}) = -\frac{1}{2} \left\{ \log 2\pi \hat{\sigma}_{s_{m'}}^2 + \frac{(s_{m't} - \hat{\mu}_{s_{m'}})^2}{\hat{\sigma}_{s_{m'}}^2} \right\}
\]

\[
\log q_m'(\lambda_n^{-1}) = (\hat{\alpha}_n - 1) \log \lambda_n^{-1} - (\hat{\beta}_n \lambda_n)^{-1} - \hat{\alpha}_n \log \hat{\beta}_n - \log \Gamma(\hat{\alpha}_n)
\]
Reversible Jump MCMC Algorithm:

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 $m = m_{\text{min}}$, then perform BIRTH step.
   3. If $m = m_{\text{max}}$, then perform DEATH step.
   4. If $m_{\text{min}} < m < m_{\text{max}}$, 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:

1. Draw samples from proposal distributions $q_{m+1}$ in $C_{m+1}$ 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.
The DEATH step is described as follows:

1. Draw samples from proposal distributions \( q_{m-1} \) in \( C_{m-1} \) subspace.
2. Compute the 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.
Experiments

- Synthetic Image Data
Experiments

- Synthetic Image Data
- Raman Spectra Data
Synthetic Image Data

Seven mixture images\(^1\) were generated by three matrices,

\[
\begin{pmatrix}
1.0 & 0.3 \\
0.3 & 1.0 \\
1.0 & 1.0 \\
0.3 & 0.3 \\
0.1 & 1.0 \\
1.0 & 0.1 \\
0.1 & 0.1
\end{pmatrix}
\begin{pmatrix}
1.0 & 1.0 & 1.0 \\
1.0 & 0.3 & 0.3 \\
0.3 & 1.0 & 0.3 \\
0.3 & 0.3 & 1.0 \\
1.0 & 1.0 & 0.0 \\
0.0 & 1.0 & 1.0 \\
1.0 & 0.0 & 1.0
\end{pmatrix}
\begin{pmatrix}
1.0 & 0.3 & 0.3 & 0.3 \\
0.3 & 1.0 & 0.3 & 0.3 \\
0.3 & 0.3 & 1.0 & 0.3 \\
0.3 & 0.3 & 0.3 & 1.0 \\
1.0 & 1.0 & 0.0 & 0.0 \\
0.0 & 1.0 & 1.0 & 0.0 \\
0.0 & 0.0 & 1.0 & 1.0
\end{pmatrix}
\]

\(^1\)The images were from http://www.cs.helsinki.fi/u/phoyer/NCimages.html
Synthetic Image Data
Synthetic Image Data

Mark Girolami and Mingjun Zhong

Reversible Jump MCMC for Non-Negative Matrix Factorization
Synthetic Image Data

Mixing Matrix

Mark Girolami and Mingjun Zhong
Reversible Jump MCMC for Non-Negative Matrix Factorization
Synthetic Image Data

Mixing Matrix

Non-Negative Matrix Factorization
Synthetic Image Data

Mixing Matrix

Non-Negative Matrix Factorization

How many images to be extracted?
1. Gibbs sampler converged in 500 iterations based on potential scale reduction factor $\hat{R}$ (Gelman and Rubin, 1992).
1. Gibbs sampler converged in 500 iterations based on potential scale reduction factor $\hat{R}$ (Gelman and Rubin, 1992).

2. RJMCMC converged in 20000 iterations based on the method of (Brooks and Giudici, 2000).
Result Analysis

Log marginal computed by TI, where the mixtures were generated from four images. Various noise levels were added.
Result Analysis

Log marginal computed by TI, where the mixtures were generated from four images. Various noise levels were added.

It shows that the estimated log-marginal likelihood prefers 4 and 5 components.
Result Analysis

Log marginal computed by TI, where the mixtures were generated from four images. Various noise levels were added.

It shows that the estimated log-marginal likelihood prefers 4 and 5 components. However, if we take into account the Monte Carlo sampling variability of the estimate there is no significant evidence that models with 4 \sim 7 components are different.
Posterior distribution of $m$ computed by RJMCMC.
The number of components $m$ recorded by RJMCMC.
Table: Estimated number of components by using BIC, TI and RJMCMC.

<table>
<thead>
<tr>
<th>Noise Level</th>
<th>2 components</th>
<th>3 components</th>
<th>4 components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.05</td>
<td>0.01</td>
<td>0.001</td>
</tr>
<tr>
<td>BIC</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>TI</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>RJMCMC</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Mark Girolami and Mingjun Zhong
Reversible Jump MCMC for Non-Negative Matrix Factorization
For example the images could be extracted after the number of images were inferred. Here is an example.

Results
This novel data set was generated from multiplexed detection of six dye labelled oligonucleotides using Surface-Enhanced Resonance Raman Scattering (SERRS) (Graham et al, 2006). 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 peaks of individual spectrum may overlap in the mixture spectra.
Multiplexed Raman Spectra

The peaks of individual spectrum may overlap in the mixture spectra.

Therefore in order to discover the target DNA sequences which were labeled by different dyes, we use NMF to extract these individual spectra from the mixtures.
We recorded 64 multiplexed Raman spectra, and each contains 574 points. So if we represent the Raman spectra as $\mathbf{X} \in \mathbb{R}_{+}^{64 \times 574}$, we then decompose it into several individual spectra. In order to recover those component spectra, we firstly need to estimate the number of components $M$. 
1. Gibbs sampler converged in 20000 iterations based on potential scale reduction factor $\hat{R}$ (Gelman and Rubin, 1992).
Convergence Diagnose

1. Gibbs sampler converged in 20000 iterations based on potential scale reduction factor $\hat{R}$ (Gelman and Rubin, 1992).

Convergence of RJMCMC is achieved when each pair of these plotted variables approach the same value.
The estimated log marginal likelihood using BIC.
Thermodynamic integration Results

The estimated log marginal likelihood using TI.

![Graph showing log marginal likelihood vs. number of components]

Mark Girolami and Mingjun Zhong
Reversible Jump MCMC for Non-Negative Matrix Factorization
Posterior distribution using RJMCMC.
RJMCMC Results

Number of components recorded by RJMMC.
Recovered Spectra

(a) Cy3
(b) FAM
(c) HEX
(d) ROX
(e) TAMRA
(f) TET

(g) ROX + TAMRA

(h) (i) (j) (k) (l)

(m) (n) (o) (p) (q) (r)

(s) (t) (u) (v) (w) (x) (y)

Noise
Conclusions

(a) RJMCMC and Thermodynamic Integration have been demonstrated to be capable of inferring the number of components, which are suitable for model selection for NMF.

(b) The BIC criterion has difficulties in identifying this number. We posit that this is due to the asymptotic approximation which underlies this criterion.
References


This work is 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.