Pseudo-marginal Markov Chain Monte Carlo for Nonnegative Matrix Factorization

Noname manuscript No.
(will be inserted by the editor)

Pseudo-marginal Markov chain Monte Carlo for nonnegative matrix factorization

Junfu Du · **Mingjun Zhong**

Received: date / Accepted: date

Abstract A pseudo-marginal Markov chain Monte Carlo (PMCMC) method is proposed for nonnegative matrix factorization (NMF). The sampler jointly simulates the joint posterior distribution for the nonnegative matrices and the matrix dimensions which indicate the number of the nonnegative components in the NMF model. We show that the PMCMC sampler is a generalization of a version of the reversible jump Markov chain Monte Carlo (RJMCMC). An illustrative synthetic data was used to demonstrate the ability of the proposed PMCMC sampler in inferring the nonnegative matrices and as well as the matrix dimensions. The proposed sampler was also applied to a nuclear magnetic resonance (NMR) spectroscopy data to infer the number of nonnegative components.

Keywords Pseudo-marginal Markov Chain Monte Carlo · Nonnegative Matrix Factorization · Reversible Jump Markov Chain Monte Carlo · Importance Sampling

1 Introduction

Given a data matrix $\mathbf{X} \in \mathbb{R}^{N\times D}$, the NMF problem is to represent $\mathbf{X}$ as a product of two unknown nonnegative matrices $\mathbf{U} \in \mathbb{R}_+^{N\times M}$ and $\mathbf{V}^T \in \mathbb{R}_+^{M\times D}$, where T denotes the transpose of a matrix, plus a noise matrix $\mathbf{E} \in \mathbb{R}^{N\times D}$, which can be conveniently represented as the following model:

$$\mathbf{X} = \mathbf{U}\mathbf{V}^T + \mathbf{E} \tag{1}$$

Note that M, N, and D are positive integers, and in this paper we assume that $M \leq N < D$. Particularly, the column of the noise matrix follows a Gaussian distribution with zero mean

Junfu Du
School of Science
Dalian Ocean University
Dalian, 116023, P.R.China
E-mail: djf@dlou.edu.cn

Mingjun Zhong (Corresponding Author)
Department of Biomedical Engineering
Dalian University of Technology
Dalian, 116024, P.R.China
E-mail: zhong@dlut.edu.cn

and an unknown diagonal covariance matrix $\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_N)$. In this model, the data matrix $\mathbf{X}$ is not necessary nonnegative. For recent approaches to NMF, please see [12]; NMF is also an important approach to dimensionality reduction [3, 19]. It is well known that the matrix dimension M (i.e., the number of nonnegative components) is unknown; even if M is known, without imposing constrains on the model, the solution is not unique. In this paper, we aim to infer M, $\mathbf{U}$, $\mathbf{V}$ and Λ simultaneously by devising a PMCMC sampler.

Given M, Gibbs samplers have been proposed for sampling the matrices $\mathbf{U}$, $\mathbf{V}$ and the noise variances λ_n [20, 16]. However, it is a difficult task to estimate M, which is essentially a model selection problem. For estimating M, the key task is to compute the analytically intractable integration $p(\mathbf{X}|M) = \int p(\mathbf{X}|\theta, M)p(\theta|M)d\theta$ where $\theta = \{\mathbf{U}, \mathbf{V}, \Lambda\}$. Thermodynamic integration (TI) [8] and Chib method [5] have been applied to compute this integration [16, 21]. It has been shown that Chib method often has numerical problems in computing some conditional densities which is required for estimating the posterior densities [21]. The TI method [8], which employs a series of power posteriors, needs to choose a suitable discretization for the temperature parameter for numerically estimating an integration; as the number of power posteriors increases, the computational cost increases. Also the discretization for the temperature parameter may affect the estimation results [4]. Reversible jump Markov chain Monte Carlo (RJMCMC) [9] has been developed to sampling M and θ for NMF simultaneously [21]. Other methods which are not sampling-based methods had also been developed for estimating M [18, 15, 17]. In this paper, a pseudo-marginal Markov chain Monte Carlo sampler (PMCMC) [2] is proposed to simulate both M and θ. We will show that the RJMCMC algorithm can be viewed as a special case of the PMCMC sampler. The proposed sampler was then applied to a synthetic data and a nuclear magnetic resonance spectroscopy data.

2 The Gibbs sampler

Prior to describe the PMCMC sampler, the Gibbs sampler is required to be derived for sampling θ for a fixed M. The Gibbs sampler will be used to generate the importance densities for approximating the marginal likelihood which is used for the PMCMC sampler. The model (1) can be represented as the form $\mathbf{X} = \sum_{m=1}^{M} \mathbf{u}_m \mathbf{v}_m^T + \mathbf{E}$ which is useful for deriving the Gibbs sampler. Denote $\mathbf{Z} = \mathbf{X} - \sum_{m=1}^{M} \mathbf{u}_m \mathbf{v}_m^T$, the likelihood for the model has the form

$$\begin{aligned} p(\mathbf{X}|\mathbf{U}, \mathbf{V}, \Lambda) &\propto \prod_{n=1}^{N} \lambda_n^{-\frac{D}{2}} \exp\left\{ -\frac{1}{2} trace\left[\mathbf{Z}^T \Lambda^{-1} \mathbf{Z}\right] \right\} \\ &\propto \prod_{n=1}^{N} \lambda_n^{-\frac{D}{2}} \exp\left\{ \mathbf{v}_m^T \widetilde{\mathbf{X}}_{-m}^T \Lambda^{-1} \mathbf{u}_m - \frac{1}{2} \mathbf{u}_m^T \Lambda^{-1} \mathbf{u}_m \mathbf{v}_m^T \mathbf{v}_m \right\} \end{aligned} \tag{2}$$

where $\widetilde{\mathbf{X}}_{-m} = \mathbf{X} - \sum_{j \neq m} \mathbf{u}_j \mathbf{v}_j^T$. This form of the likelihood is convenient to derive the conditionals for the Gibbs sampler. Since the elements of $\mathbf{U}$ and $\mathbf{V}$ are nonnegative, truncated priors are imposed on them. The u_{nm} is assumed to follow a truncated Exponential distribution such that $p(u_{nm}) = \frac{1}{1-e^{-b_u}} e^{-u_{nm}} 1_{[0,b_u]}(u_{nm})$ where $1_{[0,b_u]}(u_{nm})$ denotes that $u_{nm} \in [0, b_u]$, the λ_n^{-1} follows a Gamma prior with the form $p(\lambda_n^{-1}) = \frac{1}{\beta_\lambda^{\alpha_\lambda} \Gamma(\alpha_\lambda)} \left(\frac{1}{\lambda_n}\right)^{\alpha_\lambda - 1} e^{-\frac{1}{\beta_\lambda \lambda_n}} 1_{[0,\infty)}(\lambda_n)$, and the v_{mt} follows an Uniform prior in the range $[0, b_v]$. It is required to simulate the posterior distribution $p(\mathbf{U}, \mathbf{V}, \Lambda|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{U}, \mathbf{V}, \Lambda)p(\mathbf{U})p(\mathbf{V})p(\Lambda)$. Based on the likelihood and prior distributions, the conditionals for the Gibbs sampler have the same form with those

derived in [16] and [21].

The conditional distribution of $\mathbf{u}_m$ is a truncated Gaussian (TG) with the form

$$p(\mathbf{u}_m|\mathbf{X},\mathbf{V},\Lambda) = \mathscr{TG}(\mu_{u_m},\Sigma_{u_m},0,b_u)$$

where $\mu_{u_m} = (\mu_{u_{1m}},\cdots,\mu_{u_{Nm}})^T$ and $\Sigma_{u_m} = diag(\sigma^2_{u_{1m}},\cdots,\sigma^2_{u_{Nm}})$, where $\mu_{u_{nm}} = A^{-1}_{u_{nm}}(B_{u_{nm}} - 1)$ and $\sigma^2_{u_{nm}} = A^{-1}_{u_{nm}}$, and $\mathscr{TG}(\mu,\sigma^2,a,b)$ denotes a truncated Gaussian density with parameters μ and σ^2 defined in the range $[a,b]$. The derivation shows that $A_{u_{nm}} = \lambda_n^{-1}\mathbf{v}_m^T\mathbf{v}_m$ and $B_{u_{nm}} = \lambda_n^{-1}\mathbf{v}_m^T\widetilde{\mathbf{x}}^n_{-m}$ where $\widetilde{\mathbf{x}}^n_{-m}$ denotes the nth row of $\widetilde{\mathbf{X}}_{-m}$.

The conditional distribution of $\mathbf{v}_m$ is also a truncated Gaussian, which has the form

$$p(\mathbf{v}_m|\mathbf{X},\mathbf{U},\Lambda) = \mathscr{TG}(\mu_{v_m},\Sigma_{v_m},0,b_v)$$

where $\mu_{v_m} = A^{-1}_{v_m}B^T_{v_m}$ and $\Sigma_{v_m} = A^{-1}_{v_m}$, where $A_{v_m} = a_{v_m}I$ where $a_{v_m} = \mathbf{u}_m^T\Lambda^{-1}\mathbf{u}_m$ and $B_{v_m} = \widetilde{\mathbf{X}}^T_{-m}\Lambda^{-1}\mathbf{u}_m$.

The conditional distribution of λ_n^{-1} is a Gamma distribution,

$$p(\lambda_n^{-1}|\mathbf{x}_n,\mathbf{V},\mathbf{u}_n) = gamma(\alpha_n,\beta_n)$$

where $\alpha_n = \alpha_\lambda + D/2$ and $\beta_n = \left\{\beta_\lambda^{-1} + \frac{1}{2}\sum_{d=1}^D(x_{nd} - \mathbf{u}_n\mathbf{v}_d^T)^2\right\}^{-1}$ where $\mathbf{u}_n$ represents the nth row of $\mathbf{U}$ and $\mathbf{v}_d$ represents the dth row of $\mathbf{V}$.

The Gibbs sampler will be employed to generate the importance densities used in the pseudo-marginal Markov chain Monte Carlo method, which is proposed in the following section.

3 The pseudo-marginal Markov chain Monte Carlo sampler

It has been shown that, given M, the θ_M can be efficiently simulated using the Gibbs sampler, where θ_M represents all the parameters when the dimension is M. In this section we propose a PMCMC sampler for sampling the posterior distribution $p(M,\theta_M|\mathbf{X})$. It is straightforward to set up the proposal distribution for the Metropolis-Hastings (MH) algorithm with the form $q(M',\theta_{M'}|M,\theta_M) = q(M'|M)p(\theta_{M'}|\mathbf{X},M')$. Then the MH acceptance ratio is given by

$$\frac{p(M',\theta_{M'}|\mathbf{X})q(M,\theta_M|M',\theta_{M'})}{p(M,\theta_M|\mathbf{X})q(M',\theta_{M'}|M,\theta_M)} = \frac{p(\mathbf{X}|M')p(M')q(M|M')}{p(\mathbf{X}|M)p(M)q(M'|M)}$$

where we have used the identity $p(M,\theta_M|\mathbf{X}) = p(\theta_M|\mathbf{X},M)p(M|\mathbf{X})$. This is exactly the Bayes factor. This means that it is required to know the marginal likelihood for calculating the acceptance ratio. However the marginal likelihood is essentially the quantity we want to know. This acceptance ratio implies that the MH algorithm targets the posterior distribution $p(M|\mathbf{X})$, which is exactly what we want to infer. In most of the situations the marginal likelihood has no analytical form. Interestingly, when the marginal likelihood is not known, [2] and [1] have proposed to substitute the unknown marginal likelihood by an estimated one to compute the MH acceptance ratio, and it has been proved that under weak assumptions the algorithm leaves the target distribution $p(M,\theta_M|\mathbf{X})$ invariant. The PMCMC algorithm described in Algorithm 1 is employed to simulate the posterior distribution $p(M,\theta_M|\mathbf{X})$ for

Algorithm 1 The pseudo-marginal Markov chain Monte Carlo sampler

Input: data $\mathbf{X}$, $\alpha_\lambda = 1e-6$, $\beta_\lambda = 1e6$, $b_u = \max(\mathbf{X})$, $b_v = \max(\mathbf{X})$ and the number of samples $NSamples$.
Initialization:

- Randomly select M.
- Run the Gibbs sampler targeting the density $p(\theta_M|X,M)$ and generate the importance density denoted by $q(\theta_M|X,M)$. A sample $\theta_M(0)$ is generated by using the sampling importance resampling (SIR) technique.
- Estimate the marginal likelihood $p(\mathbf{X}|M)$ and denote the estimated marginal likelihood by $\widehat{Z}_M$.

for $i = 1$ **to** $NSamples$ **do**

- Sample M' from $q(M'|M)$.
- Run the Gibbs sampler targeting the density $p(\theta_{M'}|X,M')$ and generate the importance density denoted by $q(\theta_{M'}|X,M')$. A sample $\theta_{M'}(i)$ is generated by using the SIR technique.
- Use the importance sampling to estimate the marginal likelihood $p(\mathbf{X}|M')$ and denote the estimated marginal likelihood by $\widehat{Z}_{M'}$.
- Calculate the probability $\alpha(M,M') = \min\left\{1, \frac{\widehat{Z}_{M'}p(M')q(M|M')}{\widehat{Z}_M p(M)q(M'|M)}\right\}$.
- with probability $\alpha(M,M')$, accept $(M', \theta_{M'}(i))$, and otherwise keep $(M, \theta_M(i))$.

end for

the NMF model. The prior for M is assumed to be Uniform, and the proposal $q(M'|M)$ is also Uniform which indicates that the probabilities of moving from M to other states are equal.

To estimate the marginal likelihood, the importance sampling could be employed. Suppose $q(\theta_M|X,M)$ is the importance density. The marginal likelihood can be represented as

$$p(\mathbf{X}|M) = \int \frac{p(\mathbf{X}|\theta_M,M)p(\theta_M|M)}{q(\theta_M|X,M)} q(\theta_M|X,M)d\theta_M$$

This integration can be computed by using the Monte Carlo estimate. Suppose $\theta_M^l \sim q(\theta_M|X,M)$, the estimated marginal likelihood is thus $\widehat{Z}_M = \frac{1}{L}\sum_l w_M^l$, where the weights are

$$w_M^l = \frac{p(X|\theta_M^l,M)p(\theta_M^l|M)}{q(\theta_M^l|X,M)}$$

It is crucial to select the importance density, which is now generated in the following section.

3.1 The importance density and the sampling importance resampling technique

The importance densities were generated by using the Gibbs sampler. The posterior densities for both $\mathbf{U}$ and $\mathbf{V}$ are truncated Normal distributions. The posterior for λ_n^{-1} is a Gamma distribution. Therefore, the importance densities for u_{nm}, v_{mt} and λ_n^{-1} are truncated Normal and Gamma such that $q(u_{nm}|\mathbf{X},M) = \mathscr{TN}_{u_{nm}}(\mu_{u_{nm}}, \sigma^2_{u_{nm}}, 0, b_u)$, $q(v_{nm}|\mathbf{X},M) = \mathscr{TN}_{v_{nm}}(\mu_{v_{nm}}, \sigma^2_{v_{nm}}, 0, b_v)$ and $q(\lambda_n^{-1}|\mathbf{X},M) = Gamma(\lambda_n, \beta_n)$, where the parameters of those densities are posterior sample estimates given by the Gibbs sampler.

To generate a sample using the sampling importance resampling (SIR) technique [14], in the first step N samples $\{\theta_M^n\}_{n=1}^N$ are generated from the proposal densities $q(\theta_M|X,M)$. An approximation to the posterior distribution $p(\theta_M|X,M)$ can be represented as

$$\widetilde{p}(d\theta_M|\mathbf{X},M) = \sum_{n=1}^{N} W_M^n I_{\theta_M^n}(d\theta_M)$$

where $W_M^n = \frac{w_M^n}{\sum_{n=1}^N w_M^n}$ are called the importance weights and $I_{\theta_M^n}(\theta_M^n)$ equals 1 if the condition holds and otherwise 0. A sample is then drawn from the distribution $\widetilde{p}(d\theta_M|\mathbf{X},M)$, which is approximately distributed according to the posterior $p(\theta_M|\mathbf{X},M)$. Note that the importance sampling has also been applied to classification problems [11].

4 Relation to the reversible jump MCMC

A reversible jump MCMC algorithm could be easily derived from the proposed PMCMC sampler. When using the importance sampling to estimate the marginal likelihood, suppose just one sample was used to calculate the importance weights. The estimated marginal likelihood thus has the simple form $\widehat{Z}_{M'} = \frac{p(X|\theta^*_{M'},M')p(\theta^*_{M'}|M')}{q(\theta^*_{M'}|X,M')}$ where $\theta^*_{M'} \sim q(\theta_{M'}|X,M')$. Suppose we are proposing the move from M to M', then the acceptance ratio can be written as

$$\frac{p(X|\theta^*_{M'},M')p(\theta^*_{M'}|M')p(M')q(\theta_{M'}|X,M')q(M|M')}{p(X|\theta_M,M)p(\theta_M|M)p(M)q(\theta^*_{M'}|X,M')q(M'|M)}$$

This form of the acceptance ratio is exactly the same as the one proposed in [21] for the RJMCMC scheme. This implies that the RJMCMC algorithm is a special case of the proposed PMCMC sampler.

5 Simulation results

In this section, the PMCMC sampler is evaluated by applying it to a toy data set and a nuclear magnetic resonance (NMR) spectroscopy data. The PMCMC sampler was used to infer M, $\mathbf{U}$ and $\mathbf{V}$ simultaneously.

In the implementation of the Gibbs sampler, four free parameters b_u, b_v, α_λ and β_λ are required to be defined. We set $b_u = b_v$ to be the maximum value of the observation matrix $\mathbf{X}$, and set $\alpha_\lambda = 1e-6$ and $\beta_\lambda = 1e6$ for the prior of the inverse of the noise variance λ_n. The Gibbs sampler was used to generate the importance densities for the PMCMC and RJMCMC samplers. Seven thousand samples were generated by the Gibbs sampler for $\mathbf{U}$, $\mathbf{V}$ and Λ. The first five thousand samples were used as burn-in and the last two thousand samples were used to generate the importance densities. Both methods were used to simulate the posterior distribution $p(M,\theta_M|\mathbf{X})$.

5.1 Toy Data

To generate a toy data set, a 10×3 matrix $\mathbf{U}$ was generated by using the Exponential distribution with the rate parameter $\lambda = 1$. The matrix $\mathbf{V}^T$ with size 3×100 was uniformly generated in the range $[0,1]$. The observation matrix of size 10×100 was thus generated by $\mathbf{X} = \mathbf{U}\mathbf{V}^\mathbf{T} + \mathbf{E}$, where $\mathbf{E}$ is the Gaussian noise matrix. Thus in this toy data, the true value of M is 3.

The PMCMC and RJMCMC samplers were used to sampling the posterior distribution $p(M,\theta_M|\mathbf{X})$. The most important task is to infer the number M. The PMCMC sampler did not move to other states when $M = 3$ after a number of iterations. Figure 1 shows a trace

plot of the variable M with respect to the number of iterations, which shows that the sampler converged very fast. In the stable state $M = 3$, the estimated matrix $\mathbf{V}$ was also compared to the true one. The scatter plots of the estimated and the true matrix $\mathbf{V}$ are shown in the Figure 2. This shows the ability of the sampler to infer the matrices $\mathbf{U}$ and $\mathbf{V}$. The RJMCMC sampler was also applied to sampling the posterior distribution $p(M, \theta_M|\mathbf{X})$. As expected the RJMCMC sampler also did not move to other states when $M = 3$ after a number of iterations. Figure 1 also shows a trace plot for the variable M. Those results show that both PMMH and RJMCMC samplers are consistent. One problem for both methods is that they were not mixing. The reason is that the probability of $P(M|\mathbf{X})$ is approximately one, and so that it is very hard for the samplers to move to other states. A possible approach to make the algorithms to be mixing is to employ the tempering approaches [10].

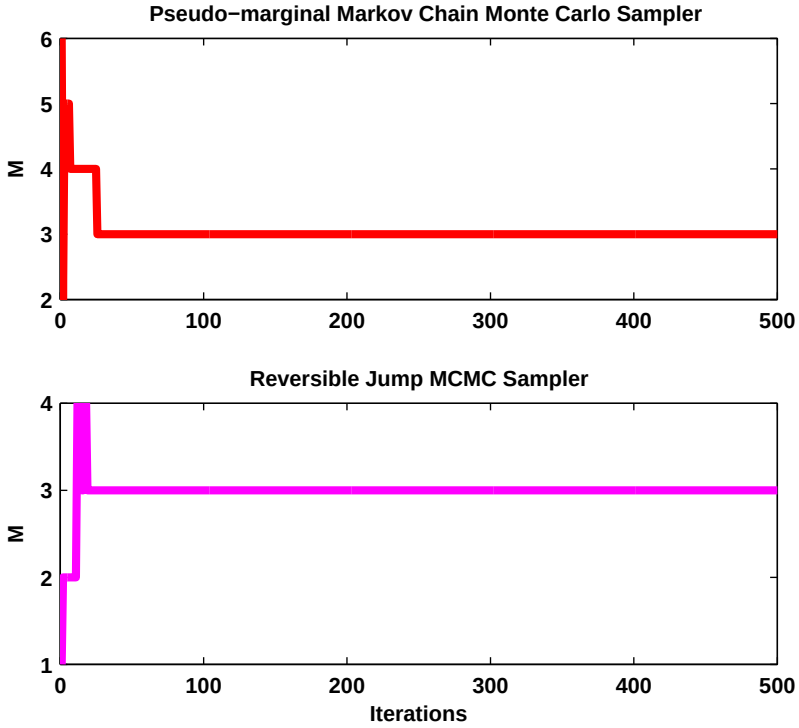

Fig. 1 The trace plot for the M in 500 iterations. Both the PMCMC and RJMCMC samplers converged to the true model state $M = 3$, quickly.

5.2 Nuclear Magnetic Resonance Spectroscopy Data

In this section the NMF model was applied to a NMR spectroscopy data which has been used to study the mixtures of metabolites in biological samplers [13]. The PMCMC sam-

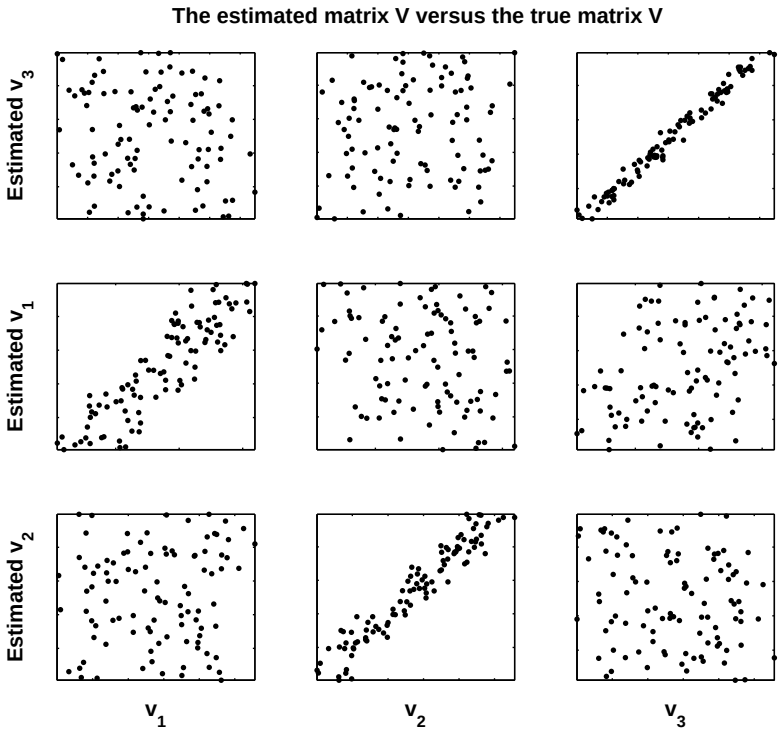

Fig. 2 The scatter plot of the estimated matrix **V** (column) versus the true matrix (row) by using the PMCMC sampler.

pler is further assessed by applying it to a data set with 8 samples acquired by *in vitro* 1D H-NMR on two neural cell types which are neurons and neural stem cells. The data were preprocessed after acquisition as usual [7], and the functional spectra is discretized by binning variables into bins of size 0.04 ppms resulting in 2394 variables totally. For each sample the baseline was removed and finally the data matrix **X** has the size 8×2394. For this data set there are also negative values in the matrix **X**. In this case, the Gibbs sampler is still valid to seek the nonnegative matrices **U** and **V**. It has been indicated by [6] that the NMF model has the character of clustering. In the NMR data, four spectroscopy samples were acquired for the neurons and the rest four samples were for neural stem cells. Therefore there should be two clusters for this NMR data. The rows of the matrix **U** should reflect the clustering results. Our task is thus to infer the **U** which indicates which sampler corresponds to which cluster, the **V** which is the collection of the intrinsic spectra for the neural types, and the M which indicates the number of clusters.

The crucial task is to infer the number of clusters, i.e., M, for the spectra data. Both the PMCMC and RJMCMC samplers were applied to seek M. As was noted both samplers were used to simulate the joint posterior distribution $p(M, \theta_M|\mathbf{X})$. Figure 3 plots the simulations for the number M using PMCMC and RJMCMC samplers. Both samplers converged

when the state moved to $M = 2$. The plots indicated that $P(M = 2|\mathbf{X}) = 1$, which correctly estimated the number of clusters. The matrices $\mathbf{U}, \mathbf{V}$ and the noise variances were also simulated along with the number M. The columns of $\mathbf{V}$ are the component spectra for neural cell types. The observation matrix $\mathbf{X}$ is the collection of the mixtures. Thus the rows of $\mathbf{U}$ could be explained as the weights for generating the mixture spectra by using the combination of the component spectra. Figure 4 shows the inferred means of the weights. The standard deviation errors are not shown, since they are relatively small. In the figure 4, one column of $\mathbf{U}$ is plotted in red and the other is in green. It shows that the matrix $\mathbf{U}$ indicated that the observation spectra $\mathbf{X}$ had two clusters, where the first four observation spectra belong to the neurons and the last four spectra belong to the neural stem cells.

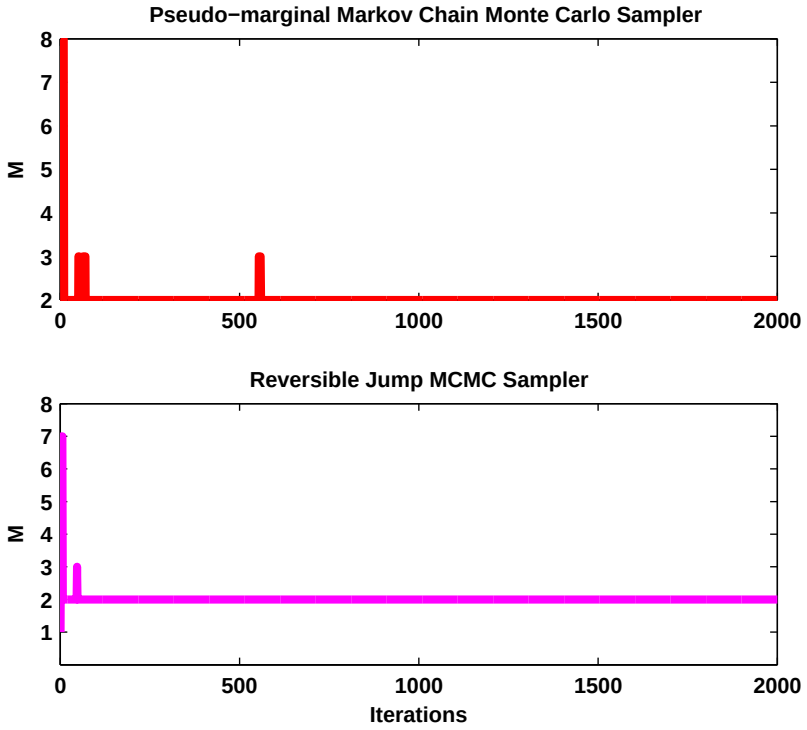

Fig. 3 The trace plot for the M in 2000 iterations. Both the PMCMC and RJMCMC samplers inferred two clusters in the data.

6 Conclusions

A pseudo-marginal Markov chain Monte Carlo method has been proposed for sampling both the matrix dimensions and the nonnegative matrices for the nonnegative matrix factorization. It has been shown that the proposed PMCMC sampler is a generalization of the RJMCMC

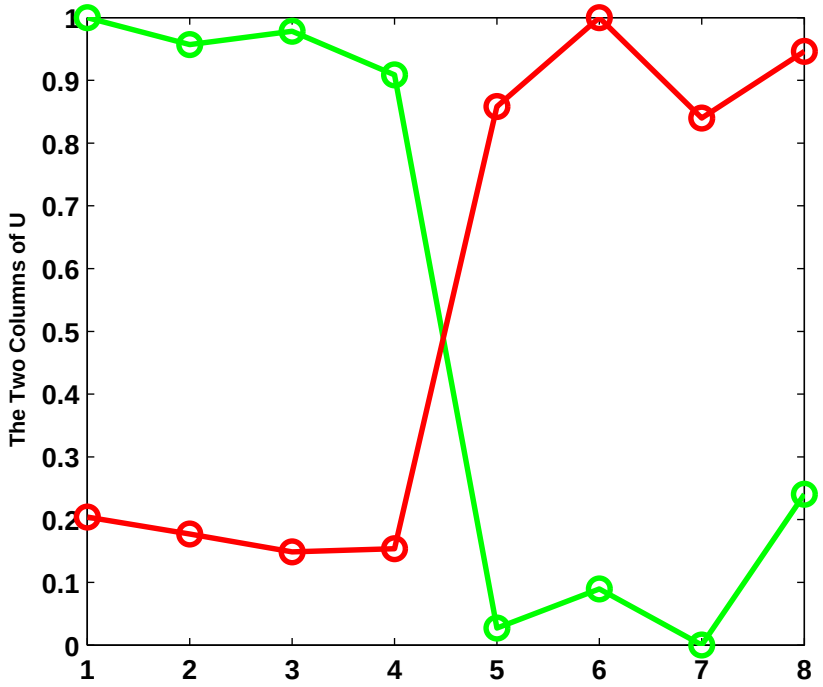

Fig. 4 The estimates of the matrix **U** by using PMCMC. The mean values (circles) of each column of **U** were divided by the maximum value of the column.

scheme employed in [21]. The PMCMC sampler was able to converge quickly and locate correctly the matrix dimensions for the toy and NMR data sets.

References

1. C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods (with discussion). *J. R. Statist. Soc. B*, 72(3):269–342, 2010.
2. C. Andrieu and G.O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.
3. Christos Boutsidis, Anastasios Zouzias, Michael W Mahoney, and Petros Drineas. Randomized dimensionality reduction for-means clustering. *Information Theory, IEEE Transactions on*, 61(2):1045–1062, 2015.
4. B. Calderhead and M. Girolami. Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics and Data Analysis*, 53:4028–4045, 2009.
5. S. Chib. Marginal likelihood from the Gibbs output. *J. Am. Statist. Ass.*, 90:1313–1321, 1995.

6. C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135, 2006.
7. W.B. Dunn, N. Bailey, and Johnson H.E. Measuring the metabolome: current analytical technologies. *Analyst*, 130:606–625, 2005.
8. N. Friel and A.N. Pettitt. Marginal likelihood estimation via power posteriors. *J. R. Statist. Soc. B*, 70:589–607, 2008.
9. P. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.
10. Jun S. Liu. *Monte Carlo strategies in scientific computing*. Springer, 2001.
11. T. Liu and D. Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis Machine Intelligence*, PP(1):1, 2015.
12. T. Liu and D. Tao. On the performance of manhattan nonnegative matrix factorization. *Neural Networks and Learning Systems, IEEE Transactions on*, PP(99):1–1, 2015.
13. L. Manganas, Zhang X, Li Y, Hazel RD, Smith SD, Wagshul ME, Henn F, Benveniste H, Djuric PM, Enikolopov G, and Maletic-Savatic M. Magnetic resonance spectroscopy identifies neural progenitor cells in the live human brain. *Science*, 318(980), 2007.
14. D. Rubin. Comment on the calculation of posterior distributions by data augmentation by Tanner, M.A. and Wong, W. H. *J. Am. Stat. Assoc.*, 82:pp. 543, 1987.
15. R Schachtner, G Po, AM Tomé, CG Puntonet, EW Lang, et al. A new bayesian approach to nonnegative matrix factorization: Uniqueness and model order selection. *Neurocomputing*, 138:142–156, 2014.
16. Mikkel N. Schmidt, Ole Winther, and Lars Kai Hansen. Bayesian non-negative matrix factorization. In *Independent Component Analysis and Signal Separation, International Conference on*, volume 5441 of *Lecture Notes in Computer Science (LNCS)*, pages 540–547. Springer, 2009.
17. Meng Sun, Xiongwei Zhang, et al. A stable approach for model order selection in nonnegative matrix factorization. *Pattern Recognition Letters*, 54:97–102, 2015.
18. V. Tan and C. Fevotte. Automatic relevance determination in nonnegative matrix factorization with the beta-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99, 2012.
19. Chang Xu, Dacheng Tao, Chao Xu, and Yong Rui. Large-margin weakly supervised dimensionality reduction. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 865–873, 2014.
20. M. Zhong and M. Girolami. Reversible jump MCMC for non-negative matrix factorization. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 663–670, Florida, USA, 2009.
21. M. Zhong, M. Girolami, K. Faulds, and D. Graham. Bayesian methods to detect dye-labelled DNA oligonucleotides in multiplexed Raman spectra. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(2):187–206, 2011.