Rank Selection for Non-negative Matrix Factorization with Normalized Maximum Likelihood Coding

Yu Ito* Shin-ichi Oeda† Kenji Yamanishi‡

Abstract

Non-negative matrix factorization (NMF) is one of the most important technologies in data mining. This is the task of factorizing a matrix into the product of two non-negative low rank matrices. In most of works on NMF, the rank is predetermined in ad hoc. This paper addresses the issue of how we can select the best rank from given data. The problem is that the conventional statistical model selection criteria such as AIC, MDL etc. cannot straightforwardly be applied to this issue because the regularity conditions for the criteria are not fulfilled. We overcome this problem to propose a novel methodology for rank selection. The key ideas are to 1) use the technique of latent variable completion to make the model regular and 2) then to apply the normalized maximum likelihood coding to rank selection for the regular model. We further propose a novel method for rank change detection when rank changes over time. We demonstrate the effectiveness of our methods for rank selection and rank change detection through synthetic data and real data sets.

1 Introduction

1.1 Motivation and Purposes of This Paper In this paper we address the following two issues: One is how to select a rank in *non-negative matrix factorization* (NMF). The other is how to detect changes of ranks in a sequential scenario. Firstly, we are concerned with the issue of selecting a rank for NMF. The task of NMF is to decompose an $N \times M$ matrix into an $N \times K$ matrix times a $K \times M$ matrix where all the elements of are non-negative. We may call K the *rank*, which is supposed to be smaller than N and M. NMF is an important task because the relation between $N \times M$ entities can be viewed through K latent factors.

In most of previous works on NMF, the rank K is assumed to be known in advance or to be given in ad hoc. In principle, however, it is desired to be determined from given data in order to explain the data best. With NMF s.t. $X = Z\Theta$, we may consider the column variable of Z and the row variable of Θ as latent variables indicating *clusters*. Rank selection is equivalent with estimation of the number of clusters, and hence is an important issue.

A critical issue is that most of statistical model selection criteria such as AIC(Akaike's information criteria) [1], BIC(Bayesian information criteria) [12], MDL(minimum description length) [10], and MML(minimum message length) [4] cannot straightforwardly be applied to rank selection. This is because NMF may be modeled using probabilistic models with latent variables, which are *irregular* in the sense that the parameters are not uniquely identifiable from data. Meanwhile, the statistical model selection criteria as above require that the models be *regular* so that the central limit theorem holds for the maximum likelihood estimator. The primary purpose of this paper is to overcome the problem to develop a new theoretically-justified methodology for rank selection for NMF

Secondly, we are concerned with the issue of *detecting rank changes*. Suppose that we are given a sequence of matrices, for each of which we would like to conduct NMF in a sequential manner. The rank for NMF may change over time. Tracking such changes is significantly important since the changes may closely be related to appearance of important events. The secondary purpose of this paper is to develop a new methodology of rank change detection.

1.2 Novelty and Significance The contribution of this paper is summarized as follows:

1) *Development of a new rank selection criterion*: We develop a new criterion for selecting the best rank for NMF. The key ideas of our criterion are summarized as follows:

1.A) Latent variable modeling of NMF. We introduce a probability model for NMF as follows: For NMF s.t. $X = Z\Theta$, the entities of X are assumed to be generated according to a probability distribution where we consider all entities of Z as latent variables and those of Θ as parameters. This latent variable modeling is inspired by the principle component analysis [2]. This probability model is irregular in the sense that there are some parameters which are not uniquely identifiable.

1.B) Normalized maximum code-length with latent variable completion. We address the rank selection issue by reducing it to the model selection of latent variable models. We should note that the latent variables are not observable in real cases. Then we may calculate the likelihood of the latent variable models by estimating the values of latent variables from the observed data. We call this technique the *latent variable completion* (LVC). Once LVC is done, the

*The University of Tokyo, dance2982002@gmail.com
†Kisarazu National College of Technology, oeda@j.kisarazu.ac.jp
‡The University of Tokyo, yamanishi@mist.i.u-tokyo.ac.jp

model would be regular and statistical model selection criteria could be applied to that model. Specifically we employ Rissanen's *minimum description length* (MDL) principle [10, 11] as a model selection criterion. It asserts that the best model is that for which the total code-length required for encoding the data as well as the model is minimum. It is here an important issue how to encode the objects. We employ the most advanced coding technique in the area of information theory, which we call the *normalized maximum likelihood* (NML) coding [11]. The reason why we use it is that the NML code-length has optimality in the sense that it attains Shtarkov's minimax criteria [13]. We give a novel formula of the NML code-length in combination with LVC.

2) Development of a novel methodology for detecting rank changes: We consider the problem of detecting rank changes for NMF from a matrix sequence. We assume that the rank probabilistically transits according to Markov chain. According to the MDL principle, we propose to sequentially select the rank so that the total sum of the code-length for data plus that for the rank transition is minimum.

3) Empirical demonstration of effectiveness of our proposed methodologies through synthetic and real data: For the synthetic data sets, we show that our proposed rank selection method is able to select the true ranks significantly more accurately than existing methods such as variational Bayes, non-parametric Bayes, straightforward application of MDL. Real data sets include image data sets and music data sets. As the evaluation metrics, we employ the prediction accuracy and AUC for the benefit vs false alarm rate curves.

1.3 Previous Works There exist a number of preceding works on rank selection for NMF. Cemgil [3] introduced a latent variable model into NMF to develop a method for rank selection using the variational Bayes method. They made a probabilistic modeling different from ours and took a Bayesian approach rather than information criteria-based one. Hoffman et al. [7] introduced a method for rank selection using the non-parametric Bayes method with Gamma process. Yamauchi et al. [16] directly applied the MDL criterion into rank selection for NMF, ignoring its irregularity. Miettinen and Vreeken [8] proposed an MDL-based approach to selecting ranks in Boolean matrix factorization (BMF). They didn't make an explicit probabilistic modeling of NMF but applied the MDL principle into BMF by considering it as a fully parametric model.

The problem of irregularity of latent variable models has been pointed out by Watanabe [14]. He took an algebraic-geometric approach to solving that problem. He developed a statistical model selection criterion called WAIC for non-identifiable models. However, WAIC is analytically difficult to calculate for NMF. The technique of latent variable completion (LVC) for information criteria has been applied to Näive Bayes model by Kontkanen and Myllymäki. [9] and to Gaussian mixture models by Hirai and Yamaishi [6].

Related to rank change detection, a general method called *dynamic model selection* (DMS) has been proposed by Yamanishi and Maruyama [15]. It is a method for selecting a sequence of statistical models on the basis of the MDL principle even when the models may change over time.

The rest of this paper is organized as follows: Section 2 proposes a method for rank selection. Section 3 proposes a method for rank change detection. Sections 4 and 5 give experimental results. Section 6 gives concluding remarks.

2 Rank Selection

2.1 Probability Model for Rank Selection We consider how to factorize a given non-negative matrix X into the product of two non-negative matrices: Z and Θ, i.e.

$$X = Z\Theta,$$

where $X \in (\mathbb{Z}^+ \cup \{0\})^{N\times M}$, $Z \in (\mathbb{R}^+ \cup \{0\})^{N\times K}$, and $\Theta \in (\mathbb{R}^+ \cup \{0\})^{K\times M}$. We call the number K the *rank* of the factorization. K is set to be smaller than N and M. We introduce the following probabilistic structure into NMF:

- X is an observed data matrix which consists of N tuples of M-dimensional vectors.
- Θ is a *parameter matrix* while Z is a *latent variable one*.
- Each entity of Z is generated according to Gamma distribution:

  $$z_{nk} \sim \mathrm{G}(z_{nk}; \alpha_{nk}, \beta_{nk}),$$

  where $\mathrm{G}(x : a, b)$ denotes the probability distribution with density function $f(x) = x^{a-1}e^{-x/b}/\Gamma(a)b^a$. A prior distribution is set to be $\mathrm{G}(z; 1/2, 2)$. This is because each entity of Z is non-negative and $\mathrm{G}(z; 1/2, 2)$ is the distribution of square of normal variables.
- Each entity of X is obtained as a sum of an intermediate variable of S so that $x_{nm} = \sum_k s_{nkm}$ and each entity of S is distributed according to Poisson distribution:

  $$s_{nkm} \sim \mathrm{Po}(s_{nkm}; z_{nk}\theta_{km}), \;\; x_{nm} = \sum_{k=1}^{K} s_{nkm}.$$

  Here $\mathrm{Po}(X; \lambda)$ denotes the probability distribution with probability mass function $f(x) = \lambda^x e^{-\lambda}/x!$. This is because the Poisson distribution is the typical distribution over $\mathbb{Z}^+ \cup \{0\}$. Although the data range is restricted to the set of integers, this setting is practically useful because the data matrix can be transformed into that of integers by multiplying sufficiently large integers to it. Note that x_{nm} also follows the Poisson distribution:

  $$x_{nm} \sim \mathrm{Po}\left(x_{nm}; \sum_k z_{nk}\theta_{km}\right).$$

We denote the distribution of X, Z, S as $P(X, Z, S; \Theta, \alpha, \beta)$.

For all (n, m), there exist $(\{z_{nk}\}, \{\theta_{km}\}) \neq (\{z'_{nk}\}, \{\theta'_{km}\})$ such that $\sum_k z_{nk}\theta_{km} = \sum_k z'_{nk}\theta'_{km}$. We say that the model is *irregular* in that case. If the model is irregular, the parameter values are not identifiable from data. This implies that the central limit theorem for the maximum likelihood estimator does not hold for such irregular models. Under such a situation, conventional statistical model selection criteria such as AIC [1], BIC [12], MDL [10], and MML [4] cannot straightforwardly be applied to rank selection, since they are derived using the central limit theorem.

2.2 Parameter Estimation for NMF When rank K is given, the issue of NMF can be reduced to the maximum likelihood estimation of Z and Θ using the EM algorithm. It follows [3]. The obtained estimates are not truly the maximum likelihood estimates but rather their approximations since the EM algorithm finds local optimal solutions. The parameter estimation algorithm is shown in Algorithm 2.1.

ALGORITHM 2.1. Parameter Estimation for NMF

Initialization: Conduct singular value decomposition for X and let the result be the initial values of $\hat{Z}$ and $\hat{\Theta}$.
Repeat the following procedure until convergence.
E-step:
Update S as follows:
$p_{nkm} = \frac{\exp(\langle \log \hat{z}_{nk} \rangle + \log \hat{\theta}_{km})}{\sum_{k=1}^{K} \exp(\langle \log \hat{z}_{nk} \rangle + \log \hat{\theta}_{km})}$,
$\hat{s}_{nkm} = \langle s_{nkm} \rangle = x_{nm} p_{nkm}$.
Update Z as follows:
$\hat{\alpha}_{nk} = \frac{1}{2} + \sum_{m=1}^{M} \langle s_{nkm} \rangle$,
$\hat{\beta}_{nk} = \left(\frac{1}{2} + \sum_{m=1}^{M} \hat{\theta}_{km} \right)^{-1}$,
$\hat{z}_{nk} = \langle z_{nk} \rangle = \hat{\alpha}_{nk} \hat{\beta}_{nk}$.
M-step:
Update Θ as follows:
$\hat{\theta}_{km} = \sum_{n=1}^{N} \langle s_{nkm} \rangle / \sum_{n=1}^{N} \langle z_{nk} \rangle$.

2.3 Rank Selection with Normalized Maximum Likelihood We introduce a criterion for selecting the best rank on the basis of the MDL principle. It is different from the existing ones in that 1) it uses the technique of *latent variable completion* (LVC) in order to make the model regular, and 2) it uses the *normalized maximum likelihood* (NML) coding to calculate the total code-length for the regular model.

Below we derive our criterion. We aim at estimating a joint distribution of X, Z, S: $P(X, Z, S; \Theta, \alpha, \beta)$ rather than the marginal distribution of X: $P(X; \Theta, \alpha, \beta)$. Note that $P(X, Z, S; \Theta, \alpha, \beta)$ is no longer an irregular model. In order to estimate the model that explains X, Z, S, we try to find the model for which the total code-length required for encoding X, Z, S is minimum, according to the MDL principle. We employ as an encoding method the *normalized maximum likelihood* (NML) code-length [11]. This is defined by the negative logarithm of the NML distribution defined as:

$$P_{\mathrm{NML}}(X, Z, S) \stackrel{\mathrm{def}}{=} \frac{P(X, Z, S; \hat{\Theta}, \hat{\alpha}, \hat{\beta})}{\mathcal{C}(K, N)},$$

where

$$\begin{aligned} &P(X, Z, S; \Theta, \alpha, \beta) \\ &\stackrel{\mathrm{def}}{=} \prod_{n=1}^{N} \prod_{m=1}^{M} \left(\delta \left(x_{nm} = \sum_{k=1}^{K} s_{nkm} \right) \right. \\ &\qquad \left. \prod_{k=1}^{K} \{ \mathrm{G}(z_{nk}; \alpha_{nk}, \beta_{nk}) \mathrm{Po}(s_{nkm}; z_{nk}\theta_{km}) \} \right), \end{aligned}$$

and $\hat{\Theta}, \hat{\alpha}, \hat{\beta}$ are respectively the maximum likelihood estimates of Θ, α, β. $\mathcal{C}(K, N)$ is the normalization term:

$$\begin{aligned} \mathcal{C}(K, N) &\stackrel{\mathrm{def}}{=} \int \sum_{X \in \mathcal{X}} \sum_{S \in \mathcal{S}} P(X, Z, S; \hat{\Theta}; \hat{\alpha}, \hat{\beta}) \mathrm{d}Z \\ &= \int \sum_{X \in \mathcal{X}} \sum_{S \in \mathcal{S}} \prod_{n=1}^{N} \prod_{m=1}^{M} \left(\delta \left(x_{nm} = \sum_{k=1}^{K} s_{nkm} \right) \right. \\ &\qquad \left. \prod_{k=1}^{K} \{ \mathrm{G}(z_{nk}; \hat{\alpha}_{nk}, \hat{\beta}_{nk}) \mathrm{Po}(s_{nkm}; z_{nk}\hat{\theta}_{km}) \} \right) \mathrm{d}Z, \end{aligned} \tag{2.1}$$

where $\mathcal{X}$ is the set of all possible Xs and $\mathcal{S}$ is the set of all possible Ss. Notice here that X is observed but neither Z nor S is observed. Then we complement Z and S from X by estimating Z and S using the EM algorithm. Letting $\hat{Z}$ and $\hat{S}$ be the estimates of Z and S respectively, we then obtain the likelihood for a regular model relative to $X, \hat{Z}, \hat{S}$. This is a step of LVC.

The next step is to combine LVC with the NML code-length. By taking the negative logarithm of the likelihood for the joint distribution with completed latent variables, the NML code-length is calculated as follows:

$$\begin{aligned} &-\log P_{\mathrm{NML}}(X, \hat{Z}, \hat{S}) \\ &= -\log P(X, \hat{Z}, \hat{S}; \hat{\Theta}, \hat{\alpha}, \hat{\beta}) + \log \mathcal{C}(K, N). \end{aligned} \tag{2.2}$$

It gives a criterion for selecting the best K. We call (2.2) the *MDL criterion*. The smaller it is, the better K is.

Since the second term in (2.2) includes the sums and integral with respect to X, Z, S, they are difficult to calculate in general. However, it can be approximated efficiently. We apply Rissanen's formula [11] for approximating the normalization term as follows: For fixed Z,

$$\begin{aligned} &\log \sum_{X \in \mathcal{X}} \sum_{S \in \mathcal{S}} P(X, S | Z; \hat{\Theta}, \hat{\alpha}, \hat{\beta}) \\ &= M \left\{ \frac{1}{2} \log \frac{N}{2\pi} + \log \int \sqrt{I(\theta_{km})} d\theta_{km} \right\}, \end{aligned} \tag{2.3}$$

where $I(\theta_{km})$ is the Fisher information, which is calculated as $1/\sqrt{z_k\theta_{km}}$ for Poisson distribution. Since for each k, each z_{nk} follows an identical distribution for any n, we denote this random variable z_k. Note that the term $\log \int \sqrt{I(\theta_{km})}d\theta_{km}$ diverges. We restrict each element in Θ to $[0, \theta_{\max}]$ for a finite $\theta_{\max}$. Since each z_k is independent, we calculate the integral with respect to Z as the product of the integral with respect to each z_k. We further restrict each component z_k of Z so they belong to $[z_{\min}, z_{\max}]$ for finite values $z_{\min}$ and $z_{\max}$. Calculating the integral with respect to Z per component, and taking the logarithm of both sides of (2.1) yield:

(2.4)

$$
\begin{aligned}
&\log \mathcal{C}(K, N) \\
&= \sum_{k=1}^{K} \log \int \exp\left[M \left\{ \frac{1}{2} \log \frac{N}{2\pi} + \log 2 + \frac{\log(z_k\theta_{\max})}{2} \right.\right. \\
&\quad \left.\left. + \log^* \left(\frac{\log(z_k\theta_{\max})}{\log 2} \right) \right\} \right] \mathrm{d}z_k \\
&= (z_{\max} - z_{\min}) MK \left(\frac{1}{2} \log \frac{N}{2\pi} + \log 2 \right) + \\
&\sum_{k=1}^{K} \log \int_{z_{\min}}^{z_{\max}} \exp\left(\frac{\log(z_k\theta_{\max})}{2} + \log^* \left(\frac{\log(z_k\theta_{\max})}{\log 2} \right) \right) \mathrm{d}z_k,
\end{aligned}
$$

where $\log^* a = \log 2.865 + \log\lfloor a \rfloor + \log\log\lfloor a \rfloor + \cdots$ and the sum is taken over all positive values. It is the code-length for $\lfloor a \rfloor$ (see [10], p.34).

We propose a strategy for selecting rank K so that the MDL criterion (2.2) with the log normalization term (2.4) is minimum. The complexity for computing the MDL criterion is $O(NKM)$. Note that $z_{\min}, z_{\max}$ and $\theta_{\max}$ play roles of hyper-parameters, which will be discussed in Section 4.2.

3 Rank Change Detection

3.1 Problem Setting Suppose that we are given a sequence of matrices. We would like to conduct NMF for each matrix. We are specifically concerned with the issue of how the rank changes over time.

First let us show how to get a sequence of matrices. Given an $N \times M$ matrix, we construct a sequence of $W \times M$ matrices sliding a window of size W. Let X_1 be a data matrix which consists of the 1st row–the Wth row in the original matrix. Similarly, let X_j be a data matrix which consists of the j-th row–the $(W + j - 1)$th row ($j = 1, \ldots, N - W + 1$). We obtain a matrix sequence $\{X_t : t = 1, \ldots, N - W + 1\}$. Here t is the time index.

Let $X_t = Z_t\Theta_t$ be NMF of X_t where $Z_t \in (\mathbb{R}^+ \cup \{0\})^{N \times K_t}$ is a latent matrix and K_t is the rank for NMF. We cannot conduct NMF for X_t independently with respect to t since Z_t and K_t may be dependent on X_j, Z_j, K_j ($j = 1, 2, \ldots, t-1$).

We are then concerned with how to sequentially estimate K_t and Z_t when given X_t and (X_j, Z_j, K_j) ($j = 1, 2, \ldots, t-1$). The process is on-line, i.e., estimation should be conducted sequentially with respect to t. If $K_t \neq K_{t-1}$, we consider that a rank change has occurred at time t.

Hereafter, for the sake of notational simplicity, we omit S_t and Θ_t. We make an assumption that K_{t-1} transits to K_t according to the following transition probability distribution:

$$
\begin{aligned}
&P(K_1 \mid K_0 : \mu) = 1/K_{\max}, \\
&P(K_t \mid K_{t-1} : \mu) = \begin{cases} 1 - \mu & (K_t = K_{t-1}), \\ \mu/2 & (K_t = K_{t-1} \pm 1), \end{cases}
\end{aligned} \tag{3.5}
$$

where $0 < \mu < 1$ is a 1-dimensional parameter and $K_{\max}$ is an upper bound on K. We make an assumption that the rank at time $t + 1$ was within ± 1 of that at time t. It is straightforward to extend it into the case where the rank at time $t + 1$ is within $\pm L$ of that at time t for any $L > 0$.

3.2 Rank Change Detection Algorithm Below we give an algorithm for rank change detection for NMF. Note that rank change detection is equivalent with sequential estimation of ranks. We conduct sequential estimation of K_t as follows: At $t = 1$ we estimate K_1 according to the procedure of Section 2, and let the selected rank be $\hat{K}_1$. At $t = 2$, we compute the total code-length for the case of $K_2 = \hat{K}_1$ and that for the cases of $K_2 = \hat{K}_1 \pm 1$. We select the rank attaining the minimum code-length among them and let it be $\hat{K}_2$. We repeat this procedure with respect to t.

We employ the framework of *dynamic model selection* (DMS)[15] for rank change detection. It selects a rank sequence so that the total code-length for the data as well as the rank transition is minimum. Below we follow Hirai and Yamanshi[5] to introduce the sequential variant of DMS.

Let $X^{t-1} = X_1 \ldots X_{t-1}$ be an observed data sequence, $Z^{t-1} = Z_1 \ldots Z_{t-1}$ be a latent variable sequence, and $K^{t-1} = K_1 \ldots K_{t-1}$be a rank sequence. We introduce the *DMS criterion* as the code-length for X_t, Z_t, K_t given $X^{t-1}, Z^{t-1}, K^{t-1}$ under the assumption K_t transits probabilistically according to (3.5). Let us denote the DMS criterion as $L(X_t, Z_t, K_t \mid X^{t-1}, Z^{t-1}, K^{t-1})$. It is given as:

$$
\begin{aligned}
&L(X_t, Z_t, K_t \mid X^{t-1}, Z^{t-1}, K^{t-1}) \\
&= l(X_t, Z_t \mid X^{t-1}, Z^{t-1} : K_t \cdot K^{t-1}) + l(K_t \mid K^{t-1}) \\
&= -\log P_{\mathrm{NML}}(X_t, Z_t; K_t) - \log P(K_t \mid K^{t-1}; \hat{\mu}_t).
\end{aligned} \tag{3.6}
$$

Here $l(X_t, Z_t \mid X^{t-1}, Z^{t-1} : K_t \cdot \hat{K}^{t-1})$ is the code-length for X_t and Z_t for given $X^{t-1}, Z^{t-1} : K_t \cdot \hat{K}^{t-1}$. It is calculated using the normalized maximum likelihood code-length of $-\log P_{\mathrm{NML}}(X_t, Z_t; K_t)$. $l(K_t \mid K^{t-1}) = -\log P(K_t|K^{t-1}; \hat{\mu}_t)$ is the code-length for K_t for given K^{t-1}. Here $\hat{\mu}_t$ is a maximum likelihood estimate of μ. We select rank K_t from $\{K_{t-1} - 1, K_{t-1}, K_{t-1} + 1\}$ that minimizes (3.6).

In the calculation of the DMS criterion, Z_t must be estimated from X_t and Z^{t-1}. Since we construct X_t from

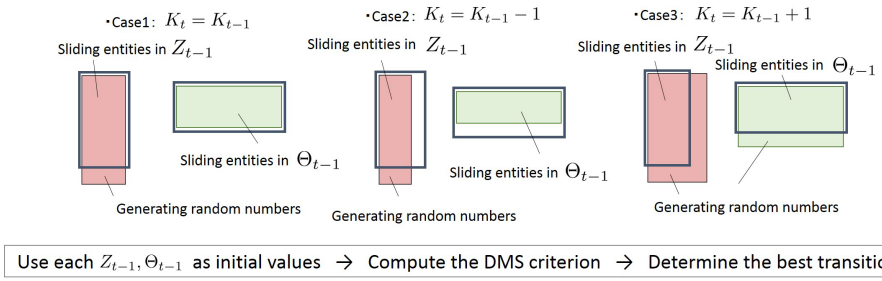


Figure 1: 3 patterns of rank changes

X_{t-1} by sliding the window, we also slide some entities in Z_{t-1} and complete new ones by generating random numbers. We consider them as initial values of Z_t, then Z_t is estimated starting from the initial values (Fig.1). The rank change detection algorithm is summarized in Algorithm 3.1.

ALGORITHM 3.1. Rank Change Detection

Initialization: Estimate Z_1 and K_1 from X_1.

for all $t = 2, \ldots, N - W + 1$ **do**

Repeat the following procedure:

Compute the DMS criterion for the following cases:

- **Case1:** $K_t = K_{t-1}$ denoted as $K_t^{(1)}$:
  Slide the entities of Z_{t-1} and complete new entities with random numbers.
  Estimate Z_t letting them be initial values. Let the obtained Z_t be $Z_t^{(1)}$.
  Calculate $L(X_t, Z_t^{(1)}, K_t \mid X^{t-1}, Z^{t-1}, K^{t-1})$, which we denote as $L^{(1)}$.
- **Case2:** $K_t = K_{t-1} - 1$ denoted as $K_t^{(2)}$:
  Slide the entities of that obtained by deleting the K_{t-1}th column from Z_{t-1} and complete new entities with random numbers.
  Estimate Z_t letting them be initial values. Let the obtained Z_t be $Z_t^{(2)}$.
  Calculate $L(X_t, Z_t^{(2)}, K_t \mid X^{t-1}, Z^{t-1}, K^{t-1})$, which we denote as $L^{(2)}$.
- **Case3:** $K_t = K_{t-1} + 1$ denoted as $K_t^{(3)}$:
  Slide the entities of that obtained by adding the $K_{t-1} + 1$th column to Z_{t-1} and complete new entities with random numbers.
  Estimate Z_t letting them be initial values. Let the obtained Z_t be $Z_t^{(3)}$.
  Calculate $L(X_t, Z_t^{(3)}, K_t \mid X^{t-1}, Z^{t-1}, K^{t-1})$, which we denote as $L^{(3)}$.

Determine the best transition that minimizes the DMS criterion (3.6).

Let $id = \underset{c}{\operatorname{argmin}} L^{(c)}$.

Update $Z_t \leftarrow Z_t^{(id)}$, $K_t \leftarrow K_t^{(id)}$

end for

Note that Z_t is estimated by letting the numerical values obtained from Z_{t-1} be initial value. This is also the case with Θ_t. Hence both Z_t and Θ_t as well as K_t are sequentially obtained with respect to t. The computational complexity of Algorithm 3.1 is $O(WNK_{\max}M)$.

4 Experiments: Rank Selection

We show experimental results on rank selection through synthetic data and real data.

4.1 Methods for Comparison We employed the following three rank selection methods for comparison:

- 1) Bayesian marginal distributions (BM),
- 2) Gamma Process NMF (GaP),
- 3) Regular MDL for latent variable models (MDL1).

All of the three methods for comparison were designed on the basis of statistical modeling of NMF. Although there may exist other possible methods for rank selection, we focus on comparison among the promising *statistical modeling-based methods* for rank selection.

BM is the method proposed by Cemgil [3]. It employs the variational Bayes method to derive the Bayesian marginal distributions for different ranks. We select the best rank so that it maximizes the Bayesian marginal distribution.

GaP is the method proposed by Hoffman et al. [7]. It employs the non-parametric Bayes method with Gamma process for modeling NMFs.

Regular MDL is a criterion obtained by applying the MDL principle into the marginal distribution (see [16]). We denote it as MDL1.

4.2 Synthetic Data Sets For various values of N, M and K, we generated synthetic data sets as follows:

- $z_{nk} \sim \mathrm{G}(1/2, 2)$,
- $\theta_{km} \sim \mathrm{Uniform}[0, 10]$,
- $x_{nm} \sim \mathrm{Po}(\sum_{k=1}^{K} z_{nk}\theta_{km})$.

We let $X \in (\mathbb{Z} \cup \{0\})^{N \times M}$ be observed variables. In order to evaluate the performance precisely, we focused on the case

where the true ranks were relatively small.

We evaluated all the methods for comparison in terms of benefit. Letting T be the total size of data, K_{true} be the true rank and $\hat{K}_j$ be the estimated rank from the j-th data, we define *benefit* as

$$B(K_{\text{true}}, \hat{K}_1, \ldots, \hat{K}_T) \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T} \max[0, 1-|K_{\text{true}}-\hat{K}_t|/U].$$

where T is the number of trials and $U = 3$. Benefit measures how close the estimated rank is to the true one. Beneft takes the maximum value 1 if and only if $\hat{K}_{\text{true}}$ coincides with the true rank K_{true}. It decreases linearly as $|K_{\text{true}} - \hat{K}_t|$ increases. If $\hat{K}_t$ is not within U of K_{true}, it is thought of as a false alarm. We set the parameters in our proposed method as: $z_{\min} = 0.8034, z_{\max} = 1.1638$, which were defined as 95% confidence interval for $\mathrm{G}(1/2, 2)$, and $\theta_{\max} = 10$ because we generated the elements of Θ according to the uniform distribution over $[0, 10]$. All of the competitive methods include a number of parameters. Through all the experiments in this paper, we tuned them so that they achieved the greatest accuracy rate.

Figs.2,3,4 show the graphs of benefit versus N for $K_{true} = 3, 5, 7$. We observe that our proposed method achieved the highest benefit values almost uniformly with respect to N in all cases. We observe that our proposed method was able to estimate ranks for NMF more precisely than the competitive ones even for smaller values of N and M. For example, in the case of $M = 30, K_{\text{true}} = 7$, for $N = 30$, our proposed method achieved more than 0.6 benefit value while all the competitive methods achieved less than 0.2 benefit values. For greater values of M and K_{true}, the differences among all the methods became small as N increased. This is because the data size became large enough for the estimation of ranks for all the methods.

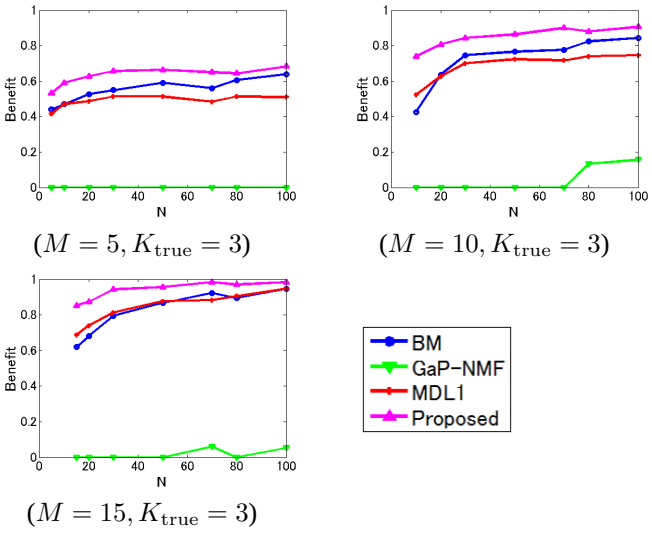
4.3 Real Data Sets

We used music and image data sets.

4.3.1 Music Signal Data Set

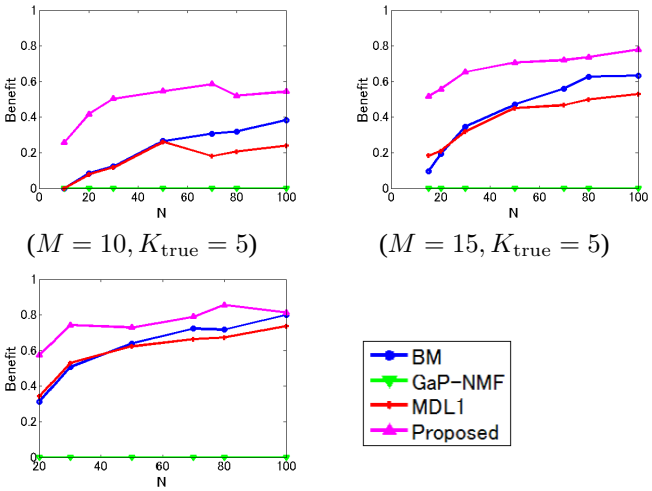
NMF has been applied to the signal analysis in music (see e.g. [7]). Rank for NMF is closely related to the number of instruments played in a piece of music. Hence the piece of music played with many instruments should be discriminated from the one played with fewer instruments in terms of rank for NMF.

We prepared 26 music signal data, each of which was represented by a spectrogram i.e., $N \times M$ matrix, where a row showed time ($N = 508 \sim 2075$) and a column showed a value of frequency ($M = 1102$). N was determined by the length of the piece while M was determined by the sampling theorem. We divided the 26 data into two groups; one was the group of of pieces played with 2 instruments (piano, organ) and the other was that of pieces played with 5 instruments (piano, guitar, bass, organ, drum) .



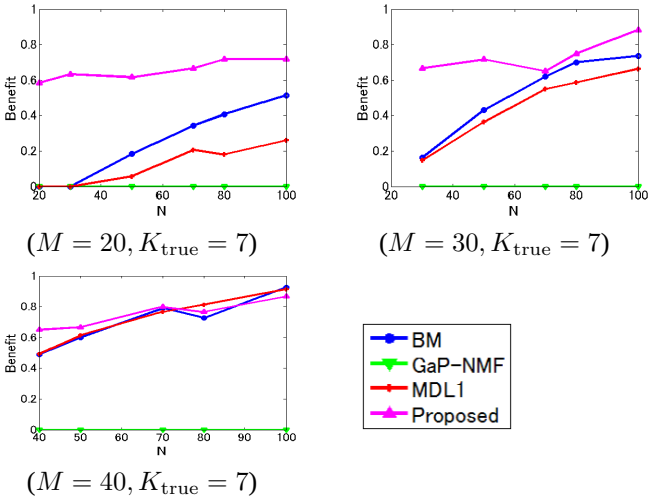
$(M = 5, K_{\text{true}} = 3)$ $(M = 10, K_{\text{true}} = 3)$ $(M = 15, K_{\text{true}} = 3)$

Figure 2: Benefit vs N for $K_{\text{true}} = 3$



$(M = 10, K_{\text{true}} = 5)$ $(M = 15, K_{\text{true}} = 5)$ $(M = 20, K_{\text{true}} = 5)$

Figure 3: Benefit vs N for $K_{\text{true}} = 5$



$(M = 20, K_{\text{true}} = 7)$ $(M = 30, K_{\text{true}} = 7)$ $(M = 40, K_{\text{true}} = 7)$

Figure 4: Benefit vs N for $K_{\text{true}} = 7$

We evaluated the effectiveness of rank selection methods by looking at how different the average of rank estimates for the first group was from that of the second one. Tables 1 and 2 show ranks selected for the first and second groups. GaP was omitted since it didn't work well in our setting as shown in the previous section. The second row shows the average of estimated ranks, and the third and fourth rows show 2.5% upper side/lower side point assuming that estimated ranks are normally distributed in the respective groups.

Table 1: Rank estimates for the first group

	BM	MDL1	Proposed
Average of estimated ranks	2.77	1.69	1.31
2.5% upper side point	5.73	3.59	2.57
2.5% lower side point	-0.19	-0.20	0.05

Table 2: Rank estimates for the second group

	BM	MDL1	Proposed
Average of estimated ranks	3.31	2.38	2.62
2.5% upper side point	4.81	3.40	3.63
2.5% lower side point	1.81	1.37	1.99

By conducting the significance difference test, it turned out that the proposed method was the only method for which the ranks estimated for the first and second ones were significantly different. It implies that our proposed method was able to appropriately estimate the ranks so that the pieces of music played with many instruments could be discriminated from that played with a few ones.

4.3.2 Image Data Sets NMF has been applied to image data sets for the purpose of completion of image data or image classification. We employed here the data set called *oliverttifaces* (http://www.cs.toronto.edu/roweis/data/olivettifaces.mat), which was used for rank estimation in [3]. This data set consists of 400 image data, each of which is represented by 64×64 pixels. We compressed each image data into data of 8×8 pixels. We consider this data set to be a matrix of size 400×64 where each row indicates the index of image and each column indicates the index of pixels. Each element shows a gray-scaled density, which takes an integer value belonging to $[0, 256]$.

We divided the matrix into a matrix of size 300×64 for training and a matrix of size 100×64 for testing. We applied BM, MDL1, and our method into the training data set to obtain the best rank and parameters. For the test data set, we made 4 missing values for each of 100 images. We conducted NMF for the test data using the learned rank and parameter values. We then employed the resulting NMF to predict values of the test data. We evaluated their prediction errors in terms of Frobenius norm.

Using the cross-validation, we repeated the process four times by selecting different training sets and calculated prediction errors of a respective method by taking an average over the four trials. Table 3 shows the results on comparison of all the methods. We see from Table 3 that our proposed method achieved the least value of prediction errors, which was significantly smaller than BM and MDL1.

Table 3: Comparison of ranks for image data set

	BM	MDL1	Proposed
Errors	506.3998	481.3368	**449.1645**

5 Experiments: Rank Change Detection

5.1 Results: Synthetic Data Sets We employed synthetic data to investigate how well our proposed method worked.

We generated the synthetic data as follows:

- Generate a matrix $X^{(1)}$ of size 300×20 with rank 2.
- Generate a matrix $X^{(2)}$ of size 200×20 with rank 3.
- Generate a matrix $X^{(3)}$ of size 200×20 with rank 5.
- Concatenate $X^{(1)}, X^{(2)}$ and $X^{(3)}$ to generate a data X of size 700×20.

The method for data generation follows Section 4.2. The rank jumps to only one higher one at the first change point, while it jumps to two higher one at the second one.

There exists no previous work on rank change detection. Hence, for the sake of comparison, we used the same framework as our proposed one, in which we replaced our rank selection method with the competitive ones as in the previous section. We examined the Bayesian marginal method (BM) and the regular MDL for latent variable model (MDL1) as in Section 4.1 for comparison. The Gamma process NMF(GaP) was out of the list since it turned out not to work as well as others as shown in Section 4.2. All of $z_{\max}, z_{\min}, \theta_{\max}$ were set as with the rank selection.

We measured the performance of rank change detection in terms of the three measures; average accuracy rate, benefit and false alarm rate. *Average accuracy rate* measures how correctly a method estimated rank. It is defined as follows:

$$\frac{1}{N}\sum_{t=1}^{N} \max\{0, 1 - |K_t^{\text{true}} - \hat{K}_t|/U_{\text{rank}}\}, \quad (N = 700),$$

where we set $U_{\text{rank}} = 3$. It takes the maximum value 1, and decreases to zero linearly as $|K_t^{\text{true}} - \hat{K}_t|$ increases with slope $1/U_{\text{rank}}$. *Benefit* measures how early a method detected the change point. It is defined as follows:

$$\frac{1}{N_{\text{CP}}}\sum_{i=1}^{N_{\text{CP}}} \max\{0, 1 - (\hat{t}_{\text{CP}} - t_i^{\text{true}})/U_{\text{time}}\}, \quad (N_{\text{CP}} = 2),$$

where we set $U_{\text{time}} = 50$ because we let the limit of permitted delay 50. t^{true} denotes the true change point and

$\hat{t}_{\rm CP}$ is the estimated change point. N_{CP} is the number of true change points. *False alarm rate* is defined as follows:

$$\frac{1}{N}\sum_{\hat{t}_{\rm CP}\in\hat{T}_{\rm CP}} \neg I(\exists t^{\rm true},\ t^{\rm true} \le \hat{t}_{\rm CP} \le t^{\rm true}+d),$$

where $\hat{T}_{\rm CP}$ is the set of all estimated change points and d is the limit of permitted delay and we set $d = 50$. $I[x]$ is a function such that $I(x) = 1$ if x is true else $I(x) = 0$ and $\neg$ denotes the negation.

Fig.5 shows how the selected ranks change over time for all the methods. We can see that our method was able to track the rank changes successfully while BM and MDL1 looked over them. Note that at the second change point the true rank changed from 3 to 5. Our method tracked that by changing rank from 3 to 4 and then from 4 to 5 step by step.

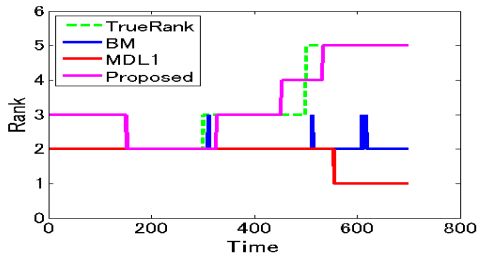


Figure 5: Result on rank change detection for synthetic data

Table 4 shows results on the comparison of our method with competitive ones in terms of accuracy rate, change detection rate, and false alarm rate, where the window size was fixed to be 100. Our proposed method achieved the highest records in terms of accuracy rate and benefit while its FAR is larger than those of the others.

Table 4: Comparison of performance for window size =100

	BM	MDL1	Proposed
Average accuracy rate	0.642	0.429	**0.839**
Benefit	0.154	0	**0.416**
False alarm rate	0.0026	**0.0014**	0.0089

We investigated how the performance depends on the window size W used in our proposed rank change detection algorithm. Fig. 6 shows graphs of benefit versus false alarm rate where each plot was obtained for a single window size ($W = 25 \sim 200$). The larger the window size was, the less false alarm rates we had. Note that the computation time depends linearly on the window size. It implies that the greater window sizes produced greater accuracy but required more computation time. Table 5 shows results on comparison of area under curve (AUC) for the benefit vs false alarm curve where benefit and false alarm rate are normalized by their maximum values over all possible window sizes. We observe that our method achieved much better performance than BM and MDL1 in terms of AUC.

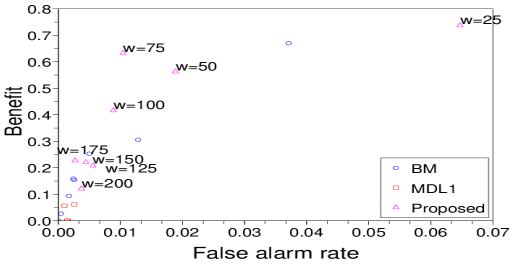


Figure 6: Relation between performance and window size

Table 5: Comparison of AUC for rank change detection

	BM	MDL1	Proposed
AUC	0.696326	0.082379	**0.787223**

5.2 Results: Real Data Sets We investigated how well our method worked for a real data set. The data set consists of spectrogram matrices for 13 pieces of music where for each matrix ($N \times M$), the row shows time ($N = 1098 \sim 4198$) and the column shows frequency ($M = 1102$).

Fig.7 shows an example of data, a matrix of size 1298×1102. Its rank was estimated as 3. Fig. 7 shows the values of each latent variable component of the three versus time. We see that the value of a certain component suddenly increased at time 646. which was thought of as the time when the number of instruments increased. We investigated whether we could detect such change points by conducting rank change detection. The change points were defined as the time points when the number of played instruments changed. We set the window size to be 200. We employed the same parameters for our method as with the synthetic data while we tuned all the parameters for the competitive methods so that they achieved the greatest accuracy rate.

Fig.8 shows how our method and the competitive ones selected ranks over time. The vertical axis shows rank while the horizontal axis shows time. The dotted vertical line shows time 646, when the number of instruments increased. The results are summarized in Table 6.

We see from Fig.8 that our method successfully detected the event at time 646 by tracking rank changes while BM and MDL1 failed to detect it. Although our method detected another change point around time 1100, it is not clear whether it is truly a false alarm or not. This is because there is some possibility that some event which caused the

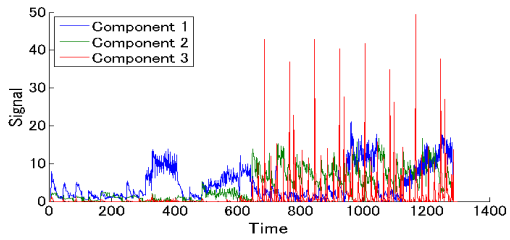


Figure 7: Example of music signal data

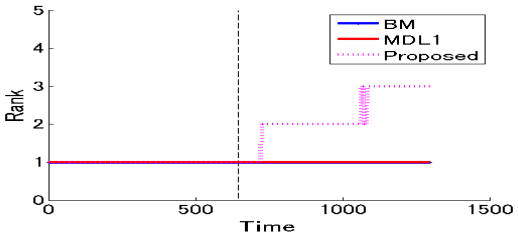


Figure 8: Result on rank change detection for real data

rank change might have happened at that point. We see from Table 6 that our method significantly outperformed BM and MDL1 with respect to average accuracy and detection rate.

In real applications it is hard to verify whether the estimated ranks are correct or not. Nevertheless, rank change points would be worthwhile being detected. This is because the change points themselves may be related to important events such as change of the number of instrumentsas in our example. Hence the rank change detection would provide a promising approach to event detection.

Table 6: Comparison for music signal data set

	BM	MDL1	Proposed
Benefit	0.2200	0	**0.7585**
False alarm rate	0.0008	**0.0002**	0.0015

6 Conclusion

We have proposed novel methodologies for rank selection and rank change detection for NMF. The rank selection method has been realized by 1) introducing a probabilistic structure with latent variables, 2) applying the technique of latent variable completion, and 3) deriving a model selection criterion based on normalized maximum likelihood coding. It can thereby overcome the difficulty which comes from the irregularity of the probabilistic structure to conduct rank selection on the basis of the MDL principle in a theoretically-justified manner. The rank change detection method has been designed by combining our rank selection method with dynamic model selection. We have empirically demonstrated using synthetic data and real data that the proposed methods worked significantly better than other statistical modeling-based ones such as the Bayesian marginal method, Gamma process method, a regular variant of MDL. The technique of latent variable completion with normalized maximum likelihood coding is not limited to rank selection for NMF but can be extended into a wide range of model selection for probabilistic models with latent variables.

Acknowledgments

This research was supported by JST-CREST, MEXT KAKENHI 23240019.

References

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

[2] C. Bishop. Variational principal components. *Proc. of ICANN99*, pages 509–514, 1999.

[3] A. T. Cemgil. Bayesian inference for nonnegative matrix factorization models. *Computational intelligence and neuroscience*, 2009.

[4] C.S.Wallace and D.M.Boulton. An information measure for classification. *Computer Journal*, 11:185–194, 1968.

[5] S. Hirai and K. Yamanishi. Detecting changes of clustering structures using normalized maximum likelihood coding. *Proc. of KDD2012*, pages 343–351, 2012.

[6] S. Hirai and K. Yamanishi. Efficient computation of normalized maximum likelihood coding for gaussian mixtures with its applications to optimal clustering. *IEEE Trans. on Inf. Theory*, 59(11):7718–7727, 2013.

[7] M. D. Hoffman, D. M. Blei, and P. R. Cook. Bayesian nonparametric matrix factorization for recorded music. *Proc. of ICML2010*, pages 439–446, 2010.

[8] P. Miettinen and J. Vreeken. Model order selection for boolean matrix factorization. *Proc. of KDD2011*, pages 51–59, 2011.

[9] P.Kontkanen and P.Myllymaki. An empirical comparison of nml clustering algorithms. *Proc. of ITDL-08*, 2008.

[10] J. Rissanen. *Stochastic Complexty in Statistical Inquiry*. World Scientific, Singapore, 1989.

[11] J. Rissanen. *Optimal Estmation of Parameters*. Cambridge, 2012.

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

[13] Y. M. Shtarkov. Universal sequential coding of single messages. *Prob. on Inf. Transmission*, 23(3):3–17, 1987.

[14] S. Watanabe. Algebraic analysis of nonidentifiable learning machines. *Neural Computation*, 13(4):899–933, 2001.

[15] K. Yamanishi and Y. Maruyama. Dynamic model selection with its applications to novelty detection. *IEEE Trans. on Inf. Theory*, 53(6):2180–2189, 2007.

[16] S. Yamauchi, M. Kawakita, and J. Takeuchi. Botnet detection based on non-negative matrix factorization and the mdl principle. *Neural Information Processing*, pages 400–409, 2012.