A New Algorithm for Multimodal Soft Coupling

Farnaz Sedighin, Massoud Babaie Zadeh, Bertrand Rivet
, and Christian Jutten *

Department of Electrical engineering, Sharif University of technology,
Tehran, Iran
GIPSA-lab, CNRS, Univ. Grenoble Alpes, Grenoble INP,
Grenoble, France
f_sedighin@ee.sharif.edu,
mbzadeh@yahoo.com,
{bertrand.rivet,christian.jutten}@gipsa-lab.grenoble-inp.fr

Abstract. In this paper, the problem of multimodal soft coupling under the Bayesian framework when the variance of the probabilistic model is unknown is investigated. Similarity of shared factors resulted from Nonnegative Matrix Factorization (NMF) of multimodal data sets is imposed in a soft manner by using a probabilistic model. In previous works, it is supposed that this probabilistic model is exactly known. However, this assumption does not always hold. In this paper it is supposed that the probabilistic model is already known but its variance is unknown. So the proposed algorithm estimates the variance of the probabilistic model along with other parameters during the factorization procedure. Simulation results with synthetic data confirm the effectiveness of the proposed algorithm.

Keywords: Nonnegative matrix factorization, Bayesian framework, Soft coupling

1 Introduction

Multimodal signals are recorded by different sensors viewing a same physical phenomenon. These signals can be of the same type (different microphones recording a same speech) or different types (audio and video recordings of a speech). Since the physical origin of the multimodal signals are the same, some similarities and correlations are expected among them. Utilizing this similarity by joint analyzing the multimodal signals is known as data fusion [1, 2]. Coupled factorization of multimodal data sets is a common approach for data fusion [3]. Coupled factorization of data sets can be achieved by coupled matrix factorization [4], coupled matrix-tensor factorization [2] or coupled tensor factorization [5].

Factorization of a 2-way array data set (matrix $\mathbf{V}^m$) can be achieved by using Nonnegative Matrix Factorization (NMF). NMF is decomposing a data

* This work has been partly supported by the European project ERC-2012-AdG-320684-CHESS.

matrix with nonnegative elements as a product of two matrices with nonnegative elements as [6]

$$\mathbf{V}^m = \mathbf{W}^m \mathbf{H}^m, \tag{1}$$

where $\mathbf{V}^m \in \Re^{F\times N}$ is the data matrix, $\mathbf{W}^m \in \Re^{F\times K}$ and $\mathbf{H}^m \in \Re^{K\times N}$ $(K < \min(F, N))$ are the factorization parameters.

Due to the correlation among multimodal datasets ($\mathbf{V}^m, m = 1, ..., M$), one or some of their factorization parameters is (are) similar to each other which is (are) named as shared factor(s). Other parameters which are particular for each data set are called unshared factors [5, 7]. Since NMF decomposition of a data set is not unique, joint (coupled) factorization of multimodal datasets and utilizing the similarity of the shared factors among them can improve the quality of factorization, and especially can reduce the indeterminacies.

In some algorithms such as [8] shared factors are assumed to be equal among the datasets. These algorithms are usually named as hard coupling algorithms. The "equality" of the shared factors is relaxed to their "similarity" in papers such as [4]. These algorithms are known as soft coupling algorithms. Soft coupling of the shared factors is studied in different papers and has different applications such as source separation [4] or speaker diarization [9]. Similarity of the shared factors is usually imposed by using penalty terms. The penalty terms can be ℓ_1 or ℓ_2 norms [4] or can be achieved in the Bayesian framework and based on the joint distribution of the shared factors [7].

The soft coupling in the Bayesian framework is proposed in [7] and is based on the statistical dependence of the shared factors which is assumed to be known. But this assumption does not always hold. Statistical dependence between the shared factors can be unknown. Even if the kind of the statistical dependence is known, parameters such as variance can be unknown. In this paper, soft coupling of the shared factors in the Bayesian framework when the variance of the statistical model is unknown is studied. Factorization parameters of a dataset are computed by the help of the parameters of another dataset using soft coupling. It is supposed that the kind of the statistical model between the shared factors (Gaussian) is known, but the variance of the model is unknown. So the variance is also estimated along with other parameters. In this paper, the update rules for updating the parameters are derived by using Majorization Minimization algorithm and exploiting auxiliary functions and an stopping criteria for stopping the update of the variance is also defined.

The paper is organized as follows. Soft coupling for NMF is reviewed in Section 2. The proposed algorithm is presented in Section 3, and finally Section 4 devoted to the experimental results.

2 Soft coupling for NMF

2.1 NMF model

As mentioned in the introduction part, NMF is decomposing a matrix $\mathbf{V}$ with nonnegative elements to the product of two matrices $\mathbf{W}$ and $\mathbf{H}$ with nonnegative elements. The decomposition is achieved by solving [6]

$$\min_{\mathbf{W}\geq 0,\mathbf{H}\geq 0} D(\mathbf{V}\|\mathbf{WH}), \tag{2}$$

where D measures the difference between $\mathbf{V}$ and $\mathbf{WH}$. Different functions are used for D such as Kulback-Leibler divergence or Itakura Saito divergence [8, 4]. Itakura-Saito divergence is defined as [8]

$$D_{\mathrm{IS}}(\mathbf{V}\|\mathbf{WH}) = \sum_{i,j} \frac{v(i,j)}{\sum_k w(i,k)h(k,j)} - \log \frac{v(i,j)}{\sum_k w(i,k)h(k,j)} - 1, \tag{3}$$

where $v(i,j)$, $w(i,k)$ and $h(k,j)$ are elements of $\mathbf{V}$, $\mathbf{W}$ and $\mathbf{H}$, respectively.

The parameters $\mathbf{W}$ and $\mathbf{H}$ in (2) are estimated during an update procedure. Multiplicative update rules with nonnegative initialization which preserve the nonnegativity property of the elements of the final parameters are proposed for estimating $\mathbf{W}$ and $\mathbf{H}$ in different papers [8, 4, 10] as

$$w(i,j) \leftarrow w(i,j) \times \frac{\sum_k h(j,k)v(i,k)/\hat{v}^2(i,k)}{\sum_k h(j,k)/\hat{v}(i,k)}, \tag{4}$$

$$h(i,j) \leftarrow h(i,j) \times \frac{\sum_k w(k,i)v(k,j)/\hat{v}^2(k,j)}{\sum_k w(k,i)/\hat{v}(k,j)}, \tag{5}$$

where $\hat{v}(i,j)$ is the (i,j)-th element of $\hat{\mathbf{V}} = \mathbf{WH}$, and $w(i,j)$ and $h(i,j)$ are the elements of $\mathbf{W}$ and $\mathbf{H}$, respectively.

2.2 Coupled NMF

Coupled factorization As mentioned in the introduction part, coupled factorization of multimodal datasets is a common approach for data fusion. Coupled factorization of two multimodal datasets in a hard manner (hard coupling) is modeled as [8]

$$\min_{\mathbf{W}_1,\mathbf{W}_2,\mathbf{H}} \lambda_1 D(\mathbf{V}_1\|\mathbf{W}_1\mathbf{H}) + \lambda_2 D(\mathbf{V}_2\|\mathbf{W}_2\mathbf{H}), \tag{6}$$

where $\mathbf{V}_1$ and $\mathbf{V}_2$ are the multimodal datasets, $\mathbf{H}$ is the shared factor, $\mathbf{W}_1$ and $\mathbf{W}_2$ are the unshared factors, and λ_1 and λ_2 are the weights of each term. For coupling in a soft manner (soft coupling) the above cost function changes to [4]

$$\min_{\mathbf{W}_1,\mathbf{W}_2,\mathbf{H}_1,\mathbf{H}_2} \lambda_1 D(\mathbf{V}_1\|\mathbf{W}_1\mathbf{H}_1) + \lambda_2 D(\mathbf{V}_2\|\mathbf{W}_2\mathbf{H}_2) + \lambda_3 \ell_p(\mathbf{H}_1,\mathbf{H}_2), \tag{7}$$

where $\mathbf{H}_1$ and $\mathbf{H}_2$ are the shared factors, $\ell_p(\mathbf{H}_1,\mathbf{H}_2)$ is the penalty term which imposes the similarity of the shared factors, and λ_3 weights the penalty term. As mentioned before, the penalty term can be for example ℓ_1 or ℓ_2 norms or can be obtained in the Bayesian framework which will be discussed in the next subsection.

Soft coupling in the Bayesian framework The problem of estimating $\mathbf{W}$ and $\mathbf{H}$ given $\mathbf{V}$ can be modeled as a Maximum A Posteriori (MAP) estimation of the parameters as [8, 7]

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}, \mathbf{V}) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\{-\log p(\mathbf{V}|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta})\}, \tag{8}$$

where $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ and p stands for the probability density function. Joint estimation of the parameters of the two multimodal datasets $\mathbf{V}_1$ and $\mathbf{V}_2$ (joint factorization of $\mathbf{V}_1$ and $\mathbf{V}_2$) can also be modeled as [7]

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}, \mathbf{V}_1, \mathbf{V}_2) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\{-\log p(\mathbf{V}_1|\boldsymbol{\theta}_1) - \log p(\mathbf{V}_2|\boldsymbol{\theta}_2) - \log p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\}, \tag{9}$$

where $\boldsymbol{\theta} = \{\mathbf{W}_1, \mathbf{H}_1, \mathbf{W}_2, \mathbf{H}_2\}$, $\boldsymbol{\theta}_1 = \{\mathbf{W}_1, \mathbf{H}_1\}$ and $\boldsymbol{\theta}_2 = \{\mathbf{W}_2, \mathbf{H}_2\}$. The third term, $\log p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, is the log of the joint density of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. In (9) it is assumed that the data sets $\mathbf{V}_1$ and $\mathbf{V}_2$ are conditionally independent given $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. $\mathbf{H}_1$ and $\mathbf{H}_2$ are the shared factors and $\mathbf{W}_1$ and $\mathbf{W}_2$ are the unshared factors.

Similar to [7], it is assumed that $\mathbf{H}_1$ is random but $\mathbf{H}_2$, $\mathbf{W}_1$ and $\mathbf{W}_2$ are deterministic , and $\mathbf{H}_1$ only depends on $\mathbf{H}_2$ (shared factors). So the last term of (9) can be written as

$$-\log p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = -\log p(\mathbf{H}_1|\mathbf{H}_2). \tag{10}$$

So the coupled factorization of two datasets with soft coupling based on Bayesian framework is modeled as [7]

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\{-\log p(\mathbf{V}_1|\boldsymbol{\theta}_1) - \log p(\mathbf{V}_2|\boldsymbol{\theta}_2) - \log p(\mathbf{H}_1|\mathbf{H}_2)\}, \tag{11}$$

where $-\log p(\mathbf{H}_1|\mathbf{H}_2)$ is the soft coupling term. In [7] it is assumed that the coupling model ($-\log p(\mathbf{H}_1|\mathbf{H}_2)$) and its parameters are known. In this paper it is assumed that although the statistical model between the shared factors are known, the variance of the model is unknown. So the variance should also be estimated along with the other parameters. This will be discussed in the next section.

3 The Proposed algorithm

As mentioned in the previous section, $-\log p(\mathbf{H}_1|\mathbf{H}_2)$ is the soft coupling term that relates the shared factors $\mathbf{H}_1$ and $\mathbf{H}_2$ in a soft manner. Supposing that p is the Gaussian probability density function and

$$(h_1(i,j)|h_2(i,j)) \perp\!\!\!\perp (h_1(i',j')|h_2(i',j')), \quad (i,j) \neq (i',j')$$

where $\perp\!\!\!\perp$ shows the independence between two random variables, and $h_1(i,j)$ and $h_2(i,j)$ are the (i,j)-th elements of $\mathbf{H}_1$ and $\mathbf{H}_2$, respectively. So the soft coupling term can be written as

$$-\log p(\mathbf{H}_1|\mathbf{H}_2) = \frac{\sum_{i,j}\|h_1(i,j)-h_2(i,j)\|^2}{2\sigma^2} + \sum_{i,j}\{\frac{1}{2}\log 2\pi + \log\sigma\}. \quad (12)$$

In (12), $h_1(i,j)$ and $h_2(i,j)$ are the (i,j)-th components of $\mathbf{H}_1$ and $\mathbf{H}_2$, respectively, σ is the model variance which is unknown. Since σ should be updated during the update procedure the last term of (12) cannot be ignored. In the above model, σ is the same for all of the elements of $\mathbf{H}_1$ and $\mathbf{H}_2$, but the problem can also be investigated when each element has a particular variance.

For modeling $-\log p(\mathbf{V}_1|\boldsymbol{\theta}_1)$, following [8], we assume that $\mathbf{V}_1$ is the Short Time Fourier Transform (STFT) matrix of a source whose elements at discrete time "n" and frequency "f", $(v_1(f,n))$, have the complex Gaussian distribution: $v_1(f,n) \sim \mathcal{N}_c(0, \sum_k w_1(f,k)h_1(k,n))$, where $w_1(f,k)$ and $h_1(k,n)$ are the elements of $\mathbf{W}_1$ and $\mathbf{H}_1$, respectively. Under this assumption, it is shown in [8] that (details can be found in [8])

$$-\log p(\mathbf{V}_1|\boldsymbol{\theta}_1) = -\log p(\mathbf{V}_1|\mathbf{W}_1\mathbf{H}_1) = D_{\text{IS}}(|\mathbf{V}_1|^2\|\mathbf{W}_1\mathbf{H}_1) + cst. \quad (13)$$

In this paper it is assumed that the second data set, $\mathbf{V}_2$, is factorized before and $\mathbf{H}_2$ has been computed and kept constant during the updating procedure. So the second term of (11) is vanished and the problem in the Bayesian approach is factorizing $|\mathbf{V}_1|^2$ to its components $\mathbf{W}_1$ and $\mathbf{H}_1$ by the help of $\mathbf{H}_2$ which has already been computed. The problem is formulated as

$$\begin{aligned} &\underset{\mathbf{W}_1,\mathbf{H}_1}{\text{argmin}}\left\{-\log p(\mathbf{V}_1|\mathbf{W}_1\mathbf{H}_1) - \log p(\mathbf{H}_1|\mathbf{H}_2)\right\}, \\ =&\underset{\mathbf{W}_1,\mathbf{H}_1}{\text{argmin}}\left\{D_{\text{IS}}(|\mathbf{V}_1|^2\|\mathbf{W}_1\mathbf{H}_1) + \frac{\sum_{i,j}\|h_1(i,j)-h_2(i,j)\|^2}{2\sigma^2} + \sum_{i,j}\log\sigma\right\}. \end{aligned} \quad (14)$$

Since the penalty term in (14) (last two terms) do not depend on $\mathbf{W}_1$, we can use (4) for updating $\mathbf{W}_1$. But new update rules are needed for updating $\mathbf{H}_1$ as well as σ. The update rules are discussed in the following subsections.

3.1 Update rule for updating $\mathbf{H_1}$

For applying the Majorization Minimization algorithm [6, 11] for minimizing $F(h)$, an auxiliary function $G(h^t, h)$ is defined as

$$\begin{aligned} G(h^t,h) &\geq F(h), \\ G(h^t,h^t) &= F(h^t), \end{aligned} \quad (15)$$

where $G(h^t,h)$ is an auxiliary function for $F(h)$ and h^t is the amount of h in the t-th iteration where $G(h^t,h^t) = F(h^t)$. An auxiliary function has the property that $F(h)$ is nonincreasing under the following update[6]

$$h^{t+1} = \underset{h}{\text{argmin}}\, G(h^t,h).$$

It means that $F(h^{t+1}) \leq F(h^t)$. So an update rule for minimizing $F(h)$ can be achieved by using its auxiliary function (details can be found in [6]). An auxiliary function for minimizing Itakura Saito divergence of (3) with respect to $\mathbf{H}$ is proposed in [11] as

$$G(\mathbf{H}|\mathbf{H}^t) = \sum_{i,j} \frac{{h^t}^2(i,j)}{h(i,j)} \left(\sum_k w(k,i) \frac{v(k,j)}{\hat{v}^2(k,j)} \right) + h(i,j) \left(\sum_k \frac{w(k,i)}{\hat{v}(k,j)} \right) + cte, \tag{16}$$

where $\hat{v}(k,j)$ is the (k,j)-th element of $\hat{\mathbf{V}} = \mathbf{W}\mathbf{H}^t$. Since the above auxiliary function is convex with respect to $\mathbf{H}$ (noting that $h(i,j) \geq 0 \quad \forall i,j$), its minimum can be found by putting its derivative to zero and finding the resulted parameters. In this paper, the Itakura Saito divergence is coupled with a term resulted from the Gaussian coupling of the shared factors. So the resulted convex auxiliary function for minimizing the cost function (14) with respect to $\mathbf{H}_1$ is

$$G_2(\mathbf{H}_1|\mathbf{H}_1^t) = G(\mathbf{H}_1|\mathbf{H}_1^t) + \frac{\sum_{i,j} \|h_1(i,j) - h_2(i,j)\|^2}{2\sigma^2}. \tag{17}$$

The derivative of the above auxiliary function with respect to $h_1(i,j)$ is

$$-\frac{{h_1^t}^2(i,j)}{h_1^2(i,j)} \left(\sum_k w_1(k,i) \frac{v_1(k,j)}{\hat{v}_1^2(k,j)} \right) + \left(\sum_k \frac{w_1(k,i)}{\hat{v}_1(k,j)} \right) + \frac{(h_1(i,j) - h_2(i,j))}{\sigma^2}. \tag{18}$$

The above equation should be set to zero and solved with respect to $h_1(i,j)$. Denoting $a(i,j) = -{h_1^t}^2(i,j) \Big(\sum_k w_1(k,i) \frac{v_1(k,j)}{\hat{v}_1^2(k,j)} \Big)$, $b(i,j) = \Big(\sum_k \frac{w_1(k,i)}{\hat{v}_1(k,j)} \Big) - \frac{h_2(i,j))}{\sigma^2}$ and $c(i,j) = \frac{1}{\sigma^2}$, (18) changes to

$$\frac{a(i,j) + b(i,j) \times {h_1}^2(i,j) + c(i,j) \times {h_1}^3(i,j)}{{h_1}^2(i,j)}, \tag{19}$$

where $a(i,j) < 0$, $c(i,j) > 0$ and the sign of $b(i,j)$ can be changed during the update procedure. The real root of the numerator of (19) is $\frac{1}{3}\big(z(i,j) + \frac{1}{z(i,j)} - 1\big)\frac{b(i,j)}{c(i,j)}$ where $z(i,j)$ is equal to (for simplicity in the notations the (i,j) is removed in the rest of the equations)

$$z = \frac{\sqrt[3]{3\sqrt{3}\sqrt{27a^2c^4 + 4ab^3c^2} - 27ac^2 - 2b^3}}{b\sqrt[3]{2}}. \tag{20}$$

For $\sqrt{27a^2c^4 + 4ab^3c^2}$ being real, the condition $b \leq \sqrt[3]{-\frac{27}{4}ac^2}$ (noting that $a < 0$) should be established. Simple calculation shows that $b \leq \sqrt[3]{-\frac{27}{4}ac^2}$ also results in $-27ac^2 - 2b^3 > 0$. So if $b \leq \sqrt[3]{-\frac{27}{4}ac^2}$, the numerator of (20) is positive and the sign of z is the same as the sign of b. The sign of $z + \frac{1}{z} - 1$ is the same as the sign of the z and the sign of z is the same as the sign of b, therefore if the constraint

$b \leq \sqrt[3]{-\frac{27}{4}ac^2}$ holds, $\frac{1}{3}\big(z + \frac{1}{z} - 1\big)\frac{b}{c}$ is positive. So $h_1(i,j) = \frac{1}{3}\big(z + \frac{1}{z} - 1\big)\frac{b}{c}$ is the positive root of (18). For when the condition $b \leq \sqrt[3]{-\frac{27}{4}ac^2}$ does not hold, for decreasing the auxiliary function and consequently the proposed cost function, if (18) > 0 then $h_1^t(i,j)$ decreases by dividing to $1+\beta$. Otherwise $h_1^t(i,j)$ is increased by multiplying to $1+\beta$. Noting that β is a small positive constant. Based on this discussion, the update procedure of $\mathbf{H}_1$ is summarized in Algorithm 1.

Algorithm 1 Update procedure of $\mathbf{H}_1$

```
if b ≤ ∛(−(27/4)ac²) then
    h_1^{t+1}(i,j) ← (1/3)(z + 1/z − 1)(b/c)        ▷ t is the number of the iteration
else
    if (18) > 0 then
        h_1^{t+1}(i,j) ← h_1^t(i,j)/(1 + β)
    else
        h_1^{t+1}(i,j) ← h_1^t(i,j) × (1 + β)
    end if
end if
```

3.2 Update rule for updating σ

Similar to $\mathbf{H}_1$, we use auxiliary function for updating σ as

$$G(\sigma|\sigma^t) = \frac{\sum_{i,j} \|h_1(i,j) - h_2(i,j)\|^2}{2\sigma^2} + (\log\sigma^t + \frac{\sigma - \sigma^t}{\sigma^t})K \times N, \tag{21}$$

where "log" function is replaced by it's tangent [11] which is the same for all of the elements of $\mathbf{H}_1$. So the last summation in (14) changes to the product of $(\log\sigma^t + \frac{\sigma-\sigma^t}{\sigma^t})$ by $(K \times N)$, the entry number of $\mathbf{H}_1$. The cost function of (21) is convex with respect to σ and the root of its derivative with respect to σ is

$$\sigma = \sqrt[3]{\frac{\sum_{i,j} \|h_1(i,j) - h_2(i,j)\|^2 \sigma^t}{K \times N}}. \tag{22}$$

So the variance of the model is updated using (22). Updating σ without any additional constraint results in converging σ to zero and finally $\mathbf{H}_1$ will be equal to $\mathbf{H}_2$. So updating of σ should be stopped after some iterations. Here, σ is updated as long as $D_{\mathrm{IS}}(|\mathbf{V}_1|^2 \| \mathbf{W}_1^{t+1}\mathbf{H}_1^{t+1}) \leq D_{\mathrm{IS}}(|\mathbf{V}_1|^2 \| \mathbf{W}_1^t\mathbf{H}_1^t)$. where $\mathbf{W}_1^t$ and $\mathbf{H}_1^t$ are the parameters of the previous iteration and $\mathbf{W}_1^{t+1}$ and $\mathbf{H}_1^{t+1}$ are the parameters of the current iteration. $D_{\mathrm{IS}}(|\mathbf{V}_1|^2 \| \mathbf{W}_1^{t+1}\mathbf{H}_1^{t+1})$ is the cost function of (14) without the coupling penalty term in the $(t+1)$-th iteration. Excessive reduction in σ gives a significant weight to the coupling term which results in too much similarity of $\mathbf{H}_1$ and $\mathbf{H}_2$. This makes $D_{\mathrm{IS}}(|\mathbf{V}_1|^2 \| \mathbf{W}_1\mathbf{H}_1)$ to

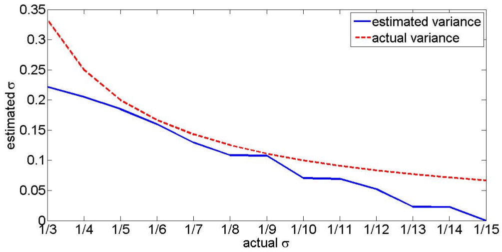

Fig. 1. Estimated variance (continuous line) using the proposed algorithm versus actual variance (dashed line).

increase (instead of decrease), especially when $\mathbf{H}_1$ and $\mathbf{H}_2$ are not very similar. This can be used as a criteria for stopping the update of σ. So updating σ stops and σ is kept fixed in the rest of the updating procedure as soon as $D_{\text{IS}}(|\mathbf{V}_1|^2\|\mathbf{W}_1^{t+1}\mathbf{H}_1^{t+1}) \leq D_{\text{IS}}(|\mathbf{V}_1|^2\|\mathbf{W}_1^{t}\mathbf{H}_1^{t})$ is violated.

4 Experimental results

In this section, the effectiveness of the proposed algorithm is investigated. In the first simulation, the quality of the proposed algorithm in estimating the variance is investigated. The matrices $\mathbf{W}_1 \in \Re^{100\times 10}$ and $\mathbf{H}_1 \in \Re^{10\times 100}$ are produced with random nonnegative elements and the data matrix ($\mathbf{V}^2$) is produced by multiplying $\mathbf{W}_1$ and $\mathbf{H}_1$. $\mathbf{H}_2$ is produced by adding Gaussian noise to $\mathbf{H}_1$ as $p(\mathbf{H}_1|\mathbf{H}_2) = \mathcal{N}(\mathbf{H}_2, \sigma)$ where $\mathcal{N}(\mathbf{H}_2, \sigma)$ is the Gaussian noise with mean of $\mathbf{H}_2$ and variance of σ which is unknown. β is set to 0.1.

The result for estimating σ is shown in Fig. 1. It is clear from the results that the algorithm has the ability to estimate the variance of the model. It is also clear that the estimated variance is decreasing by decreasing the actual variance (σ). The estimation error of $\mathbf{H}_1$ is calculated as $\|\mathbf{H}_1 - \hat{\mathbf{H}}_1\|_F^2$ where $\hat{\mathbf{H}}_1$ is the estimation of $\mathbf{H}_1$. The estimation error for the proposed algorithm and for the situation when the variance is known is presented in Fig. 2. The results show that except for some large values of σ, the proposed algorithm and the situation in which the variance is known has nearly the same estimation errors. Note that when the variance is known, only $\mathbf{W}_1$ and $\mathbf{H}_1$ are updated using (4) and Algorithm 1.

The decreasing property of the proposed cost function under the proposed update rules is shown in Fig. 3. The proposed algorithm is executed for the actual σ equal to 0.1 and $\beta = 10^{-3}$. It is clear that the cost function decreases during the update procedure.

In Table 1, the estimation error of the proposed algorithm is compared to the hard coupling algorithm. It is clear from the results that the proposed algorithm has a lower estimation error comparing to the hard coupling algorithm, especially for greater variances. But by decreasing the variance the estimation errors become closer to each other.

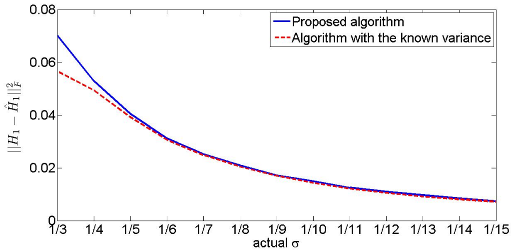

Fig. 2. Comparing the estimation errors using the proposed algorithm (continuous line) and the situation when the variance is known (dashed line).

Table 1. Estimation error of $\mathbf{H}_1$ for the proposed and hard coupling algorithms.

actual σ	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{7}$	$\frac{1}{8}$	$\frac{1}{9}$	$\frac{1}{10}$	$\frac{1}{11}$	$\frac{1}{12}$
Proposed algorithm	0.073	0.055	0.041	0.031	0.024	0.019	0.016	0.0134	0.0111	0.0099
Hard coupling	0.087	0.062	0.045	0.034	0.026	0.021	0.017	0.0139	0.0116	0.0099

Table 2. Estimation error of $\mathbf{H}_1$ for the proposed algorithm and when the variance is chosen arbitrarily.

actual σ	chosen σ					proposed algorithm
	3	1	0.3	0.1	0.03	
0.3	0.0733	0.0308	0.0481	0.0714	0.0794	0.0471
0.1	0.0731	0.0131	0.0112	0.0113	0.0121	0.0102
0	0.0769	0.00070	0.0032	1.0289×10^{-7}	5.3682×10^{-10}	3.389×10^{-21}

And finally, we have compared the proposed algorithm with the algorithm when the variance is not estimated but chosen arbitrarily (not necessarily equal to the actual variance) for several amounts of the actual σ. The estimation errors are presented in Table 2 (the estimation errors of the proposed algorithm is presented in the last column). It is clear from the results that choosing an incorrect variance especially when the actual $\sigma = 0$, can result in a significant estimation error. But this error is reduced by using the proposed algorithm.

5 Conclusion

In this paper, we have proposed an algorithm for soft coupling of the shared factors based on the Bayesian framework. As mentioned before, in soft coupling based on the Bayesian framework the statistical dependency between the shared factors should be known. But this assumption does not always hold. In this paper, it is assumed that the general statistical dependency between the shared factors (Gaussian distribution) is known but the variance of the model is unknown. So the proposed algorithm estimates the variance of the model along

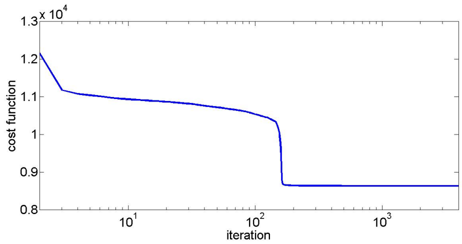

Fig. 3. Decreasing property of the proposed cost function.

with the estimation of the factorization parameters. The presented results show the ability of the proposed algorithm in estimating the model variance and also the decreasing property of the proposed algorithm.

References

1. Lahat, D., Adali, T., Jutten, C.: Challenges in multimodal data fusion. In: Proceedings of the 22nd European Signal Processing Conference (EUSIPCO), 2014. 101–105
2. Acar, E., Kolda, T.G., Dunlavy, D.M.: All-at-once optimization for coupled matrix and tensor factorizations. arXiv preprint arXiv:1105.3422 (2011)
3. Acar, E., Rasmussen, M.A., Savorani, F., Næs, T., Bro, R.: Understanding data fusion within the framework of coupled matrix and tensor factorizations. Chemometrics and Intelligent Laboratory Systems **129** (2013) 53–63
4. Seichepine, N., Essid, S., Févotte, C., Cappé, O.: Soft nonnegative matrix co-factorization. IEEE Transactions on Signal Processing **62**(22) (2014) 5940–5949
5. Rivet, B., Duda, M., Guérin-Dugué, A., Jutten, C., Comon, P.: Multimodal approach to estimate the ocular movements during eeg recordings: a coupled tensor factorization method. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE (2015) 6983–6986
6. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems. (2001) 556–562
7. Cabral Farias, R., Cohen, J.E., Comon, P.: Exploring multimodal data fusion through joint decompositions with flexible couplings. arXiv preprint arXiv:1505.07717 (2015)
8. Ozerov, A., Févotte, C.: Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Transactions on Audio, Speech, and Language Processing **18**(3) (2010) 550–563
9. Seichepine, N., Essid, S., Févotte, C., Cappé, O.: Soft nonnegative matrix co-factorizationwith application to multimodal speaker diarization. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013. 3537–3541
10. Sawada, H., Kameoka, H., Araki, S., Ueda, N.: Multichannel extensions of non-negative matrix factorization with complex-valued data. IEEE Transactions on Audio, Speech, and Language Processing **21**(5) (2013) 971–982
11. Févotte, C.: Majorization-minimization algorithm for smooth itakura-saito non-negative matrix factorization. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011. 1980–1983