

A New Algorithm for Multimodal Soft Coupling

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

▶ **To cite this version:**

Farnaz Sedighin, Massoud Babaie Zadeh, Bertrand Rivet, Christian Jutten. A New Algorithm for Multimodal Soft Coupling. LVA/ICA 2017 - 13th International Conference on Latent Variable Analysis and Signal Separation, Olivier Michel; Nadège Thirion-Moreau, Feb 2017, Grenoble, France. pp.162 - 171, 10.1007/978-3-319-53547-0_16 . hal-01479306

HAL Id: hal-01479306
https://hal.science/hal-01479306

Submitted on 28 Feb 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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 the joint analysis of the multimodal signals is known as data fusion [1, 2]. Coupled factorization of the multimodal data sets is a common approach for data fusion [3] and can be achieved by coupled matrix factorization [4], coupled matrix-tensor factorization [2] or coupled tensor factorization [5].

Factorization of matrix $\mathbf{V}^m$ (a 2-way array data set) 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]

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

where $\mathbf{V}^m \in \Re_+^{F \times N}$ is the$m$-th data set, $\mathbf{W}^m \in \Re_+^{F \times K}$ and $\mathbf{H}^m \in \Re_+^{K \times N}$ ($K < \min(F, N)$) are the factorization parameters of the m-th data set and M is the number of the data sets.

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

In some algorithms such as [8] the shared factors are assumed to be equal among the data sets. These algorithms are usually named as hard coupling algorithms. The "equality" constraint of the shared factors is relaxed to their "similarity" in algorithms such as [4]. These algorithms are known as soft coupling algorithms and are exploited in different applications such as source separation [4] or speaker diarization [9]. The similarity of the shared factors is usually controlled by using penalty terms. The penalty terms can be in the form of 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 studied in [7] and is based on the statistical dependence between the shared factors which is assumed to be known. But this assumption does not always hold. The statistical dependence between the shared factors can be unknown. Even if the kind of the statistical dependence is known, its parameters such as its variance can be unknown. In this paper, the 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 data set are computed by the help of the parameters of another data set 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 the 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 is devoted to the experimental results.

2 Soft coupling for NMF

2.1 NMF model

As mentioned in the introduction, 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 the Kulback-Leibler divergence or the Itakura Saito divergence [8, 4]. The Itakura-Saito divergence is defined as [8]

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

where $v(f,n)$, $w(f,k)$ and $h(k,n)$ are the 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 of the elements of the final parameters are proposed for estimating $\mathbf{W}$ and $\mathbf{H}$ in different papers [8, 4, 10] as

$$w(f,k) \leftarrow w(f,k) \times \frac{\sum_n h(k,n)v(f,n)/\hat{v}^2(f,n)}{\sum_n h(k,n)/\hat{v}(f,n)}, \tag{4}$$

$$h(k,n) \leftarrow h(k,n) \times \frac{\sum_f w(f,k)v(f,n)/\hat{v}^2(f,n)}{\sum_f w(f,k)/\hat{v}(f,n)}, \tag{5}$$

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

2.2 Coupled NMF

Coupled factorization As mentioned in the introduction, the coupled factorization of the multimodal data sets is a common approach for data fusion. Coupled factorization of two multimodal data sets 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 data sets, $\mathbf{H}$ 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 coupled factorization 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 controls the similarity of the shared factors, and λ_3 weights the penalty term. As mentioned before, the penalty term can be for example in the form of 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 $\mathbf{S}$ can be modeled as the Maximum A Posteriori (MAP) estimation of the parameters as [8, 7]

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

where $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ and p denotes the probability density function. The joint estimation of the parameters of the two multimodal data sets $\mathbf{S}_1$ and $\mathbf{S}_2$ can also be modeled as [7]

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

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 logarithm of the joint density of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. In (9) it is assumed that the data sets $\mathbf{S}_1$ and $\mathbf{S}_2$ 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 joint estimation of the parameters in the Bayesian framework is modeled as [7]

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

where $-\log p(\mathbf{H}_1|\mathbf{H}_2)$ relates the shared factors and is the soft coupling term. For modeling $-\log p(\mathbf{S}_i|\boldsymbol{\theta}_i)(i = 1, 2)$, in [8], it is assumed that $\mathbf{S}_i$ is the Short Time Fourier Transform (STFT) matrix of a source ($F \times N$ matrix) whose elements at discrete time "n" and frequency "f", $(s_i(f, n))$, have the complex Gaussian distribution: $s_i(f, n) \sim \mathcal{N}_c(0, \sum_k w_i(f, k)h_i(k, n))$, where $w_i(f, k)$ and $h_i(k, n)$ are the elements of $\mathbf{W}_i$ and $\mathbf{H}_i$, respectively. Under this assumption, it is shown in [8] that (details can be found in [8])

$$-\log p(\mathbf{S}_i|\boldsymbol{\theta}_i) = -\log p(\mathbf{S}_i|\mathbf{W}_i\mathbf{H}_i) = D_{\mathrm{IS}}(\mathbf{V}_i^s \| \mathbf{W}_i\mathbf{H}_i) + cst, \tag{12}$$

where $\mathbf{V}_i^s \in \Re_+^{F \times N}$ is a matrix whose elements are $v_i^s(f, n) = |s_i(f, n)|^2$. 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 we assume 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

In this paper, it is assumed that the second data set, $\mathbf{V}_2^s = |\mathbf{S}_2|^2$, is factorized beforehand and $\mathbf{H}_2$ has been computed and kept constant during the updating procedure. In addition, the variance of the model is unknown and should be estimated along with the other parameters. So the problem in the Bayesian approach is factorizing $\mathbf{V}_1^s$ to its parameters $\mathbf{W}_1$ and $\mathbf{H}_1$ and computing the model variance by the help of $\mathbf{H}_2$ which has already been computed. The problem is formulated as

$$\begin{aligned} &\underset{\mathbf{W}_1,\mathbf{H}_1,\sigma}{\operatorname{argmax}}\, p(\mathbf{S}_1,\mathbf{H}_2,\mathbf{W}_1,\mathbf{H}_1,\sigma) = \\ &\underset{\mathbf{W}_1,\mathbf{H}_1,\sigma}{\operatorname{argmin}} \Big\{ -\log p(\mathbf{S}_1|\mathbf{W}_1\mathbf{H}_1) - \log p(\mathbf{H}_1|\mathbf{H}_2,\sigma) \Big\}, \end{aligned} \tag{13}$$

where σ^2 is the variance which is unknown. 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. Recall that it is assumed that $\mathbf{S}_1$ only depends on $\mathbf{W}_1$ and $\mathbf{H}_1$ and $\mathbf{W}_1$ and σ are assumed to be deterministic. Supposing that p is the Gaussian probability density function and

$$(h_1(k,n)|h_2(k,n),\sigma) \perp\!\!\!\perp (h_1(k',n')|h_2(k',n'),\sigma), \quad (k,n) \neq (k',n')$$

where $\perp\!\!\!\perp$ shows the independence between two random variables, and $h_1(k,n)$ and $h_2(k,n)$ are the (k,n)-th elements of $\mathbf{H}_1$ and $\mathbf{H}_2$, respectively. So the soft coupling term can be written as $-\log p(\mathbf{H}_1|\mathbf{H}_2,\sigma) = \frac{\sum_{k,n}\|h_1(k,n)-h_2(k,n)\|^2}{2\sigma^2} + \sum_{k,n}\{\frac{1}{2}\log 2\pi + \log\sigma\}$. By considering (12), (13) can be written as

$$\underset{\mathbf{W}_1,\mathbf{H}_1,\sigma}{\operatorname{argmin}} \left\{ D_{\mathrm{IS}}(\mathbf{V}_1^s\|\mathbf{W}_1\mathbf{H}_1) + \frac{\sum_{k,n}\|h_1(k,n)-h_2(k,n)\|^2}{2\sigma^2} + \sum_{k,n}\log\sigma \right\}. \tag{14}$$

Since the last two terms of the cost function (14) 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 $\mathbf{H}_1$

The update rule for estimating $\mathbf{H}_1$ is derived by using the majorization minimization approach [6, 11] and by the help of the auxiliary functions. For minimizing $F(h)$, an auxiliary function $G(h^t,h)$ is defined as

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

where $G(h^t,h)$ is an auxiliary function for $F(h)$ and h^t is the point that $G(h^t,h^t)$ is equal to $F(h^t)$. $G(h^t,h^t)$ has the property that $F(h)$ is nonincreasing under the following update [6]

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

It means that $F(h^{t+1}) \leq F(h^t)$. So an update rule for minimizing $F(h)$ can be achieved by using a proper 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

$$G(\mathbf{H}|\mathbf{H}^t) = \sum_{k,n} \left\{ \frac{{h^t}^2(k,n)}{h(k,n)} \sum_f w(f,k) \frac{v(f,n)}{\hat{v}^2(f,n)} + h(k,n) \sum_f \frac{w(f,k)}{\hat{v}(f,n)} \right\} + cst, \tag{16}$$

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

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

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

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

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

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

where $a(k,n) < 0$, $c(k,n) > 0$ and $b(k,n)$ can be positive or negative. One of the roots of the numerator of (19) is $\frac{1}{3}\big(z(k,n) + \frac{1}{z(k,n)} - 1\big)\frac{b(k,n)}{c(k,n)}$ where $z(k,n)$ is equal to (for simplicity in the notations, (k,n) 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 held. Simple calculation shows that if $b \leq \sqrt[3]{-\frac{27}{4}ac^2}$ then $-27ac^2 - 2b^3$ is also positive. 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 $h_1^t(k,n)$ decreases by dividing to $1+\beta$. Otherwise $h_1^t(k,n)$ is increased by multiplying to $1+\beta$ where β is a small positive constant. Based on this discussion, the update procedure of $\mathbf{H}_1$ is summarized below:

Algorithm 1 Update procedure for $\mathbf{H}_1$ ($(t+1)$-th iteration)

if $b \leq \sqrt[3]{-\frac{27}{4}ac^2}$ **then**
  $h_1^{t+1}(k,n) \leftarrow \frac{1}{3}\left(z + \frac{1}{z} - 1\right)\frac{b}{c}$
else
  if (18) > 0 **then**
    $h_1^{t+1}(k,n) \leftarrow h_1^t(k,n)/(1+\beta)$
  else
    $h_1^{t+1}(k,n) \leftarrow h_1^t(k,n) \times (1+\beta)$
  end if
end if

3.2 Update rule for updating σ

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

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

where "log" function is replaced by its 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 auxiliary function (21) is convex with respect to σ and the root of its derivative with respect to σ is

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

So σ is updated using (22). Updating σ without any additional constraint results in the convergence of σ to zero (very small values) and $\mathbf{H}_1$ will become equal to $\mathbf{H}_2$ and finally the cost function converges to $-\infty$. So updating of σ should be stopped after some iterations.In this paper, σ is updated as long as $D_{\mathrm{IS}}(\mathbf{V}_1^s \| \mathbf{W}_1^{t+1}\mathbf{H}_1^{t+1}) \leq D_{\mathrm{IS}}(\mathbf{V}_1^s \| \mathbf{W}_1^t\mathbf{H}_1^t)$, where $\mathbf{W}_1^t$ and $\mathbf{H}_1^t$ are the parameters of the t-th iteration and $\mathbf{W}_1^{t+1}$ and $\mathbf{H}_1^{t+1}$ are the parameters of the $(t+1)$-th iteration. $D_{\mathrm{IS}}(\mathbf{V}_1^s \| \mathbf{W}_1^t\mathbf{H}_1^t)$ is the cost function of (14) without the coupling penalty term in the t-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 $D_{\mathrm{IS}}(\mathbf{V}_1^s \| \mathbf{W}_1\mathbf{H}_1)$ to 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 $D_{\mathrm{IS}}(\mathbf{V}_1^s \| \mathbf{W}_1^{t+1}\mathbf{H}_1^{t+1}) \leq D_{\mathrm{IS}}(\mathbf{V}_1^s \| \mathbf{W}_1^t\mathbf{H}_1^t)$ is violated.

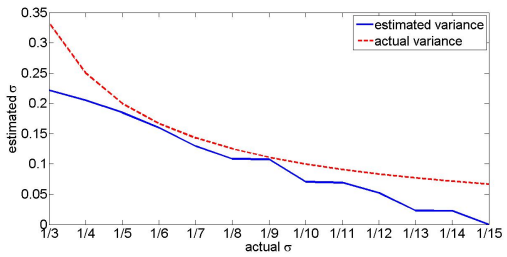


Fig. 1. The estimated σ (continuous line) using the proposed algorithm versus the actual σ (dashed line).

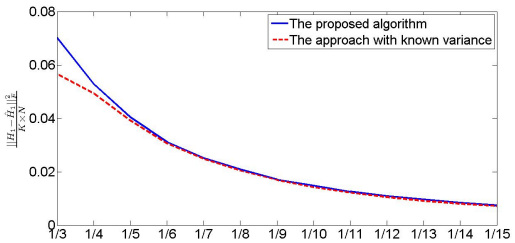


Fig. 2. Comparing the estimation errors using the proposed algorithm (continuous line) and the situation when the variance is known (dashed line).

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 $\mathbf{H}_2 \in \Re_+^{10\times 100}$ are produced with random nonnegative elements. $\mathbf{H}_1$ is produced by adding Gaussian noise to $\mathbf{H}_2$ as $p(\mathbf{H}_1|\mathbf{H}_2,\sigma) = \mathcal{N}(\mathbf{H}_2,\sigma^2)$ where $\mathcal{N}(\mathbf{H}_2,\sigma^2)$ is the Gaussian noise with the mean $\mathbf{H}_2$ and the variance σ^2. The data matrix $(\mathbf{V}_1^s)$ is produced by multiplying $\mathbf{W}_1$ and $\mathbf{H}_1$. β is set to 0.1 and all of the parameters are initialized randomly with positive values. The results for the estimation σ are shown in Fig. 1. It is clear from the results that the algorithm has the ability to estimate σ.

The estimation error of $\mathbf{H}_1$ is calculated as $\frac{\|\mathbf{H}_1-\hat{\mathbf{H}}_1\|}{K\times N}$ where $\hat{\mathbf{H}}_1$ is the estimation of $\mathbf{H}_1$. The estimation errors for the proposed algorithm and for the situation when the variance is known are shown 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 actual variance (σ^2) 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

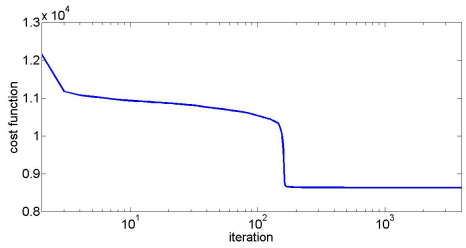


Fig. 3. Decreasing property of the proposed cost function.

Table 1. Estimation errors of $\mathbf{H}_1$ for the proposed algorithm and the hard coupling situation.

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 errors of $\mathbf{H}_1$ for the proposed algorithm and when σ 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}

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 situation in which $\hat{\mathbf{H}}_1 = \mathbf{H}_2$. It is clear from the results that the proposed algorithm has a lower estimation error comparing to the hard coupling situation, especially for greater variances. But by decreasing the variance the estimation errors become closer to each other.

And finally, we have compared the proposed algorithm with the situation when the variance is not estimated but is 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 the soft coupling of the shared factors in the Bayesian framework. As mentioned before, for the soft coupling of the shared factors in the Bayesian framework the statistical model between the shared factors should be known. But this assumption does not always hold. In this paper, it is assumed that the general statistical model 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 with the estimation of the factorization parameters. The presented results show the ability of the proposed algorithm in the estimation of the model variance and also the decreasing property of the proposed algorithm.

References

1. Lahat, D., Adal, T., Jutten, C.: Challenges in multimodal data fusion. In: 2014 22nd European Signal Processing Conference (EUSIPCO). (Sept 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 cofactorization. IEEE Transactions on Signal Processing **62**(22) (Nov 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). (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. Farias, R.C., Cohen, J.E., Comon, P.: Exploring multimodal data fusion through joint decompositions with flexible couplings. IEEE Transactions on Signal Processing **64**(18) (Sept 2016) 4830–4844
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) (March 2010) 550–563
9. Seichepine, N., Essid, S., Févotte, C., Cappé, O.: Soft nonnegative matrix cofactorizationwith application to multimodal speaker diarization. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. (May 2013) 3537–3541
10. Sawada, H., Kameoka, H., Araki, S., Ueda, N.: Multichannel extensions of nonnegative matrix factorization with complex-valued data. IEEE Transactions on Audio, Speech, and Language Processing **21**(5) (May 2013) 971–982
11. Févotte, C.: Majorization-minimization algorithm for smooth itakura-saito nonnegative matrix factorization. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (May 2011) 1980–1983