A SPARSE OPTIMIZATION APPROACH TO SUPERVISED NMF BASED ON CONVEX ANALYTIC METHOD

Yu Morikawa and Masahiro Yukawa

Dept. Electrical and Electronic Engineering, Niigata University, Japan

ABSTRACT

In this paper, we propose a novel scheme to supervised nonnegative matrix factorization (NMF). We formulate the supervised NMF as a sparse optimization problem assuming the availability of a set of basis vectors, some of which are irrelevant to a given matrix to be decomposed. The proposed scheme is presented in the context of music transcription and musical instrument recognition. In addition to the nonnegativity constraint, we introduce three regularization terms: (i) a block ℓ_1 norm to select relevant basis vectors, and (ii) a temporal-continuity term plus the popular ℓ_1 norm to estimate correct activation vectors. We present a state-of-the-art convex-analytic iterative solver which ensures global convergence. The number of basis vectors to be actively used is obtained as a consequence of optimization. Simulation results show the efficacy of the proposed scheme both in the case of perfect/imperfect basis matrices.

Index Terms— supervised nonnegative matrix factorization, sparse optimization, convex analysis

1. INTRODUCTION

Nonnegative matrix factorization (NMF) has been an active research topic of great importance in signal processing over the decades [1–6]. The problem can be stated simply as follows: decompose a given nonnegative data-matrix $\boldsymbol{Y}$ into $\boldsymbol{W}\boldsymbol{H}(\approx \boldsymbol{Y})$, where both matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ must be nonnegative. Here, $\boldsymbol{W}$ is a basis matrix (a set of basis vectors) and $\boldsymbol{H}$ is an activation matrix (a set of activation vectors). A challenging issue in the (unsupervised) NMF is the *source-number determination*; i.e., it is required to know the exact number of basis vectors prior to decomposition [7, 8]. A larger number of columns in $\boldsymbol{W}$ (than it should be) would result in producing undesired basis vectors, while a smaller number of columns would fail in capturing desired basis vectors.

The major contribution of this paper is to present a sparse optimization scheme to supervised NMF based on a state-of-the-art technique developed in convex analysis. We present the proposed scheme in the context of polyphonic music transcription and musical instrument recognition although it is applicable to general supervised NMF problems. A remarkable advantage is that there is no need to make the *source-number determination* prior to decomposition. The proposed approach is based on the assumption that a set of basis vectors, some of which are *irrelevant* to the matrix $\boldsymbol{Y}$, is available in decomposition, and the source number is determined automatically as the number of *relevant* basis vectors obtained by sparse optimization. (The terms *relevant* and *irrelevant* are explained in Section 2.) To select the relevant basis vectors and to estimate the activation vectors accurately in response to $\boldsymbol{Y}$, we adopt the sparse optimization framework with the following regularization terms, in addition to an indicator function penalty imposing the nonnegativity constraint: (i) a block ℓ_1 norm and (ii) temporal-continuity term [6] and the ℓ_1 norm of a vectorized $\boldsymbol{H}$ (i.e., the sum of the absolute values of all entries of $\boldsymbol{H}$). The block ℓ_1 norm plays a role in selecting the relevant basis vectors from the basis matrix. The temporal-continuity term together with the ℓ_1 norm contributes mainly to estimating the correct activation vectors. From the convex optimization viewpoint, the cost function can be seen as a sum of a smooth convex function (the data-fidelity term plus the temporal-continuity term) and the three nonsmooth, but *proximable*, convex functions (an indicator function to enforce nonnegativity, the block ℓ_1 norm, and the ℓ_1 norm). Here, *proximable* means that the Moreau proximity operators defined in Section 2.2 can be computed. To find a minimizer of the cost function, we present the iterative solver of *generalized forward-backward splitting* (*GFBS*) algorithm [9], which guarantees convergence to an optimal solution. The simulation results show the efficacy of the proposed scheme in an application to polyphonic music transcription and music instrument recognition.

Relation to prior work: Grindlay and Ellis have proposed a statistical approach to supervised NMF under an assumption (similar to the current study) on the availability of $\boldsymbol{W}$ [10]. Their method is based also on an additional statistical assumption that "*a suitably normalized magnitude spectrum can be modeled as a joint distribution over time and frequency*". The expectation maximization algorithm is employed therein, and hence convergence to a global solution is *not* ensured in general. Our approach is, in contrast, deterministic (relying on no statistical assumption) and, as mentioned already, global convergence is ensured. Dessein *et al.* have proposed a real-time event-detection scheme based on convex quadratic programming [11]. The approach has been developed for real-time processing and could fail in estimating activation vectors correctly even in simple numerical simulations as will be shown in Section 3, although global convergence to an optimal point of their cost function is ensured. This is because the approach makes no use of temporal information, while the proposed scheme exploits it via the temporal-continuity term and the ℓ_1 norm. Indeed, the proposed scheme is quite flexible and expandable in the sense that other possible convex penalties could be easily incorporated, although it is beyond the scope of the present work.

2. SUPERVISED NMF BY SPARSE OPTIMIZATION

We assume that the basis matrix $\boldsymbol{W}$ is given but it may contain such basis vectors that are *irrelevant* to $\boldsymbol{Y}$ as well as *relevant* ones; the terms *relevant* and *irrelevant* are explained as follows. Assume for instance that we have four basis vectors, say $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3, \boldsymbol{w}_4$, which represent notes of piano (pitches: C, D) and notes of guitar (pitches: C, D) respectively. If $\boldsymbol{Y}$ is generated from an audio signal composed of a note of piano-C ($\boldsymbol{w}_1$) and a note of guitar-D ($\boldsymbol{w}_4$), we refer to $\boldsymbol{w}_1$ and $\boldsymbol{w}_4$ as *relevant* basis vectors and to $\boldsymbol{w}_2$ and $\boldsymbol{w}_3$ as *irrelevant* ones.

To select relevant basis vectors, we formulate the supervised NMF as a sparse optimization problem. This offers an attractive way to avoid making the *source-number determination* prior to decomposition. As will be seen in Section 3, the proposed scheme is robust against possible imperfection in $\boldsymbol{W}$ (i.e., mismatches between $\boldsymbol{W}$ and the set of true basis vectors associated with the instruments present in the input audio signal). (How to learn $\boldsymbol{W}$ prior to decomposition practically is described in Section 2.2.)

2.1. Problem Formulation

We formulate the supervised NMF problem as the following sparse optimization problem (see, e.g., [12] for a comprehensive tutorial about sparse optimization):

$$(\mathrm{P}_0)\quad \min_{\boldsymbol{H}\in C} \|\boldsymbol{H}\|_{\text{row-0}} \quad \text{s.t.} \quad \|\boldsymbol{Y}-\boldsymbol{W}\boldsymbol{H}\|_{\mathrm{F}}^2 \le \epsilon_1,$$
$$\sum_{l=1}^{L}\sum_{n=1}^{N-1}(h_{l,n+1}-h_{l,n})^2 \le \epsilon_2,$$
$$\sum_{l=1}^{L}\|\hat{\boldsymbol{h}}_l\|_0 \le \epsilon_3,$$

where $\boldsymbol{Y}\in\mathbb{R}_{\ge 0}^{M\times N}$, which denotes the set of all nonnegative valued matrices of size $M\times N$, $\boldsymbol{W}\in\mathbb{R}_{\ge 0}^{M\times L}$, $\boldsymbol{H}=\begin{bmatrix}\hat{\boldsymbol{h}}_1\ \hat{\boldsymbol{h}}_2\ \cdots\ \hat{\boldsymbol{h}}_L\end{bmatrix}^{\mathsf{T}}\in\mathcal{H}:=\mathbb{R}^{L\times N}$, $h_{l,n}$ denotes the (l,n)-entry of $\boldsymbol{H}$, $C:=\mathbb{R}_{\ge 0}^{L\times N}\subset\mathcal{H}$, $\|\cdot\|_{\text{row-0}}$ the row-ℓ_0 norm which counts the number of nonzero row vectors, $\|\cdot\|_{\mathrm{F}}$ the Frobenius norm, $\|\cdot\|_0$ the ℓ_0 norm which counts the number of nonzero entries, $\epsilon_1,\epsilon_2>0$ are small constants, and ϵ_3 a positive integer. Here, $(\cdot)^{\mathsf{T}}$ stands for *transpose*, $\mathcal{H}$ is the Hilbert space of the activation matrix $\boldsymbol{H}$ to be optimized, and $\hat{\boldsymbol{h}}_l$s are referred to as activation vectors. The problem (P_0) is difficult to solve directly in practice because it has a combinatorial nature. We therefore introduce convex relaxations, reformulating (P_0) into the following unconstrained convex optimization problem:

$$(\mathrm{P}_1)\quad \min_{\boldsymbol{H}\in\mathcal{H}} J(\boldsymbol{H}) = \|\boldsymbol{Y}-\boldsymbol{W}\boldsymbol{H}\|_{\mathrm{F}}^2 + \underbrace{i_C(\boldsymbol{H})}_{\text{(a)}} + \underbrace{\lambda_1\sum_{l=1}^{L}\|\hat{\boldsymbol{h}}_l\|_2}_{\text{(b)}}$$
$$\underbrace{+\lambda_2\sum_{l=1}^{L}\sum_{n=1}^{N-1}(h_{l,n+1}-h_{l,n})^2+\lambda_3\sum_{l=1}^{L}\sum_{n=1}^{N}|h_{l,n}|}_{\text{(c)}},$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm. The penalty terms in (P_1) are detailed below.

(a) **Nonnegativity constraint:** The term $i_C(\boldsymbol{H})$ is an indicator function defined as follows:

$$i_C(\boldsymbol{H})=\begin{cases}0, & \text{if } \boldsymbol{H}\in C,\\ \infty, & \text{otherwise,}\end{cases}$$

which enforces an optimal $\boldsymbol{H}$ to be nonnegative.

(b) **Basis-vector selection:** The term $\lambda_1\sum_{l=1}^{L}\|\hat{\boldsymbol{h}}_l\|_2$ is the block ℓ_1 norm penalty for selecting the basis vectors relevant to $\boldsymbol{Y}$. Typically, the number N of the columns of $\boldsymbol{Y}$ corresponds to a few seconds and many of the basis vectors tend to be irrelevant to $\boldsymbol{Y}$, implying that many of the activation vectors would be the zero vectors (i.e., $\boldsymbol{H}$ would be block sparse).

(c) **Correct activation-vector estimation:** The term $\lambda_2\sum_{l=1}^{L}\sum_{n=1}^{N-1}(h_{l,n+1}-h_{l,n})^2$ enhances temporal continuity by suppressing the differences between the values in adjacent frames of each activation vector [6]. The term $\lambda_3\sum_{l=1}^{L}\sum_{n=1}^{N}|h_{l,n}|$ is the popular ℓ_1 norm of $\mathrm{vec}(\boldsymbol{H}):=\begin{bmatrix}\hat{\boldsymbol{h}}_1^{\mathsf{T}}\hat{\boldsymbol{h}}_2^{\mathsf{T}}\cdots\hat{\boldsymbol{h}}_L^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}}$, which is expected to discriminate active and inactive frames due to its sparsity-promoting property. The two terms jointly contribute to estimating the activation vector correctly.

2.2. Numerical Algorithm

The cost function in (P_1) can be written in the following form:

$$J(\boldsymbol{H}) = \underbrace{\varphi(\boldsymbol{H})}_{\text{smooth}} + \underbrace{\sum_{j=1}^{3}\psi_j(\boldsymbol{H})}_{\text{nonsmooth}},$$

where

$$\varphi(\boldsymbol{H}) := \|\boldsymbol{Y}-\boldsymbol{W}\boldsymbol{H}\|_{\mathrm{F}}^2+\lambda_2\sum_{l=1}^{L}\sum_{n=1}^{N-1}(h_{l,n+1}-h_{l,n})^2,$$
$$\psi_1(\boldsymbol{H}) := i_C(\boldsymbol{H}),$$
$$\psi_2(\boldsymbol{H}) := \lambda_1\sum_{l=1}^{L}\|\hat{\boldsymbol{h}}_l\|_2,$$
$$\psi_3(\boldsymbol{H}) := \lambda_3\sum_{l=1}^{L}\sum_{n=1}^{N}|h_{l,n}|.$$

Here, φ is a differentiable convex function with the Lipschitz-continuous gradient (i.e., φ is *smooth*) while ψ_j, $j=1,2,3$, are nonsmooth but *proximable* convex functions. (See [13, 14] for details about convex analysis in Hilbert spaces.)

Definition 1 *In the real Hilbert space $(\mathcal{H},\|\cdot\|_{\mathrm{F}})$, we define the following.*

(a) *Given any proper and lower-semicontinuous convex function $\psi:\mathcal{H}\to\mathbb{R}$, the proximity operator of ψ of index $\gamma>0$ for any $\boldsymbol{X}\in\mathcal{H}$ is defined as*

$$\mathrm{prox}_{\gamma\psi}(\boldsymbol{X}) := \operatorname*{argmin}_{\boldsymbol{Y}\in\mathcal{H}}\left(\psi(\boldsymbol{Y})+\frac{1}{2\gamma}\|\boldsymbol{X}-\boldsymbol{Y}\|_{\mathrm{F}}^2\right),$$

(b) *Given any nonempty closed convex set $K\subset\mathcal{H}$, the metric projection of any $\boldsymbol{X}\in\mathcal{H}$ onto the set K is defined as*

$$P_K(\boldsymbol{X}) := \operatorname*{argmin}_{\boldsymbol{Y}\in K}\|\boldsymbol{X}-\boldsymbol{Y}\|_{\mathrm{F}}.$$

Note that the proximity operator is a generalization of the metric projection because

$$\mathrm{prox}_{\gamma i_K} = P_K,\ \forall\gamma>0.$$

The problem (P_1) can iteratively be solved by generating the sequence of the auxiliary variables $(\boldsymbol{Z}_j^{(k)})_{k\in\mathbb{N}}\subset\mathcal{H}$, $j=1,2,3$, and $(\boldsymbol{H}^{(k)})_{k\in\mathbb{N}}\subset\mathcal{H}$, with initial estimates $\boldsymbol{Z}_j^{(0)}$, $j=1,2,3$, and $\boldsymbol{H}^{(0)}:=\sum_{j=1}^{3}\omega_j\boldsymbol{Z}_j^{(0)}$ as follows:

$$\boldsymbol{Z}_j^{(k)} := \boldsymbol{Z}_j^{(k-1)}+\alpha\Big(\mathrm{prox}_{\frac{\gamma}{\omega_j}\psi_j}(2\boldsymbol{H}^{(k-1)}-\boldsymbol{Z}_j^{(k-1)}$$
$$-\gamma\nabla\varphi(\boldsymbol{H}^{(k-1)}))-\boldsymbol{H}^{(k-1)}\Big),\ \ j=1,2,3,$$
$$\boldsymbol{H}^{(k)} := \sum_{j=1}^{3}\omega_j\boldsymbol{Z}_j^{(k)},$$

where

$$\omega_j \in (0,1) \ \text{ s.t. } \sum_{j=1}^{3} \omega_j = 1, \quad j = 1,2,3,$$

$$\alpha \in \left(0, \min\left\{\frac{3}{2}, \frac{\eta\gamma+2}{2\eta\gamma}\right\}\right), \tag{1}$$

$$\gamma \in \left(0, \frac{2}{\eta}\right), \tag{2}$$

$$\eta = 2\sigma_{\max}(\bar{\boldsymbol{W}}^{\mathsf{T}}\bar{\boldsymbol{W}} + \lambda_2\bar{\boldsymbol{D}}),$$

$$\bar{\boldsymbol{W}} = \begin{bmatrix} \boldsymbol{W} & & \boldsymbol{O} \\ & \boldsymbol{W} & \\ & \ddots & \\ \boldsymbol{O} & & \boldsymbol{W} \end{bmatrix} \in \mathbb{R}^{NM\times NL},$$

$$\bar{\boldsymbol{D}} = \boldsymbol{D} \otimes \boldsymbol{I}_L \in \mathbb{R}^{NL\times NL},$$

$$\boldsymbol{D} = \sum_{n=1}^{N-1} (\boldsymbol{e}_{n+1} - \boldsymbol{e}_n)(\boldsymbol{e}_{n+1} - \boldsymbol{e}_n)^{\mathsf{T}} \in \mathbb{R}^{N\times N},$$

where $\{\boldsymbol{e}_n\}_{n=1}^N$ denotes the standard basis for $\mathbb{R}^N$, $\otimes$ denotes the Kronecker product [15], $\boldsymbol{I}_L$ denotes the $L \times L$ identity matrix, and $\sigma_{\max}(\bar{\boldsymbol{W}}^{\mathsf{T}}\bar{\boldsymbol{W}} + \lambda_2\bar{\boldsymbol{D}})$ is the maximum modulus of the eigenvalues of $\bar{\boldsymbol{W}}^{\mathsf{T}}\bar{\boldsymbol{W}} + \lambda_2\bar{\boldsymbol{D}}$.

The gradient $\nabla\varphi$ and the proximity operators $\mathrm{prox}_{\frac{\gamma}{\omega_2}\psi_2}$ and $\mathrm{prox}_{\frac{\gamma}{\omega_3}\psi_3}$ can be computed as follows:

$$\nabla\varphi(\boldsymbol{H}) = 2\boldsymbol{W}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{H} - 2\boldsymbol{W}^{\mathsf{T}}\boldsymbol{Y} + 2\lambda_2\boldsymbol{H}\boldsymbol{D},$$

$$\mathrm{prox}_{\frac{\gamma}{\omega_2}\psi_2}(\boldsymbol{H}) = \sum_{l=1}^{L} \max\left\{1 - \frac{\lambda_1\gamma}{\omega_2\|\hat{\boldsymbol{h}}_l\|_2}, 0\right\} \boldsymbol{e}_l\hat{\boldsymbol{h}}_l^{\mathsf{T}},$$

$$\mathrm{prox}_{\frac{\gamma}{\omega_3}\psi_3}(\boldsymbol{H}) = \sum_{l=1}^{L}\sum_{n=1}^{N} \mathrm{sgn}(h_{l,n}) \max\left\{|h_{l,n}| - \frac{\lambda_3\gamma}{\omega_3}, 0\right\} \boldsymbol{E}_{l,n},$$

where sgn denotes the signum function and $\boldsymbol{E}_{l,n}$ is the $L\times N$ matrix having one at the (l,n)-entry and zeros elsewhere.

2.3. Discussion on Proposed Scheme

The proposed scheme ensures the global convergence as stated below.

Theorem 1 *The sequence $(\boldsymbol{H}^{(k)})_{k\in\mathbb{N}}$ generated by the algorithm presented above is convergent to a minimizer of $J(\boldsymbol{H})$, which is a minimizer of $\widetilde{J}(\boldsymbol{H}) := \varphi(\boldsymbol{H}) + \psi_2(\boldsymbol{H}) + \psi_3(\boldsymbol{H})$ over the nonnegativity constraint set C.*

Proof: It is readily verified by [9, Theorem 2.1]. □

It is seen from (1) and (2) that the upper bounds of the step size parameters α and γ depend on η which is computationally expensive to obtain because the size of $\bar{\boldsymbol{W}}$ is typically large. The following lemma is useful in practice to reduce computational costs.

Lemma 1

$$\sigma_{\max}(\bar{\boldsymbol{W}}^{\mathsf{T}}\bar{\boldsymbol{W}} + \lambda_2\bar{\boldsymbol{D}}) \le \sigma_{\max}(\bar{\boldsymbol{W}}^{\mathsf{T}}\bar{\boldsymbol{W}}) + \sigma_{\max}(\lambda_2\bar{\boldsymbol{D}}) = \sigma_{\max}(\boldsymbol{W}^{\mathsf{T}}\boldsymbol{W}) + \sigma_{\max}(\lambda_2\bar{\boldsymbol{D}}).$$

Proof: It can be verified by Rayleigh-Ritz theorem [16]. □

Lemma 1 implies that, instead of η, one may use $\tilde{\eta} := 2\sigma_{\max}(\boldsymbol{W}^{\mathsf{T}}\boldsymbol{W}) + 2\sigma_{\max}(\lambda_2\bar{\boldsymbol{D}}) \ge \eta$ which is computationally less expensive and offers step size parameters no greater than the respective upper bounds. We therefore use $\alpha := (1-\varepsilon_\alpha)\min\{3/2, (\tilde{\eta}\gamma+2)/2\tilde{\eta}\gamma\}$ and $\gamma := (2-\varepsilon_\gamma)/\tilde{\eta}$ for small constants $\varepsilon_\alpha := 1.0\times10^{-4}$, $\varepsilon_\gamma := 2.0\times10^{-4}$.

Remark 1

(a) *How to learn $\boldsymbol{W}$: Each column vector of $\boldsymbol{W}$ is generated simply by unsupervised NMF. Various single isolated tones generated from a wide variety of instruments are used as training audio signals, and unsupervised NMF is performed for each tone which is assumed to be associated with a sole basis vector. In a string of this implementation over all tones, basis vectors corresponding to the isolated tones are acquired eventually, and normalizing those basis vectors yields $\boldsymbol{W}$. The single isolated tones for this process are available at, e.g., RWC music database* [17]*, MAPS database* [18]*, etc.*

(b) *Using the Moore-Penrose pseudo-inverse: One may think that the Moore-Penrose pseudo-inverse $\boldsymbol{W}^{\dagger}$ of $\boldsymbol{W}$ could be used for computing $\boldsymbol{H}$ as $\boldsymbol{H} = \boldsymbol{W}^{\dagger}\boldsymbol{Y}$ when $M > L$ and rank($\boldsymbol{W}$) $= L$, which is often the case in the context of the present study. In our experiments, however, this approach always resulted in unsuccessful decomposition; $\boldsymbol{H}$ tends to contain negative entries and to also be a dense matrix.*

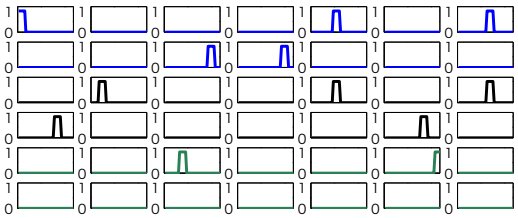
3. SIMULATION RESULTS

3.1. Case of Perfect $\boldsymbol{W}$

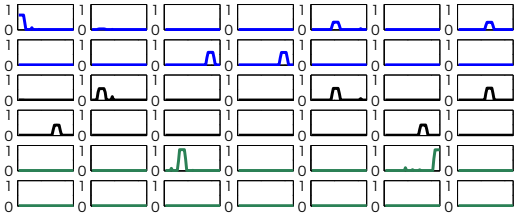
We show the efficacy of the proposed scheme for polyphonic music transcription and musical instrument recognition in the case of 'perfect' $\boldsymbol{W}$; i.e., we assume that the basis vectors to represent the input magnitude spectrogram $\boldsymbol{Y}$ are exactly known. As the timbre of the input audio signal, we use the following three types of wave: sine waves, triangle waves, and square waves. The waves are generated by matlab. For simplicity, only 14 tones of the waves are used. The basis matrix $\boldsymbol{W}$ is composed of amplitude spectra which are respectively obtained by the short-time Fourier transform (STFT) of the three waves of the same 14 tones. This is the case of perfect $\boldsymbol{W}$ since the spectra of these waves are time-invariant.

We compare the performance of the proposed scheme with the Beta Non-negative Decomposition (BND) method [11], since it has been reported that the BND method performs better than the other approaches and ensures the global convergence to the minimizer of its cost function although it has been proposed primarily for real-time processing [11]. All the audio signals for both the input audio signal and the audio signals to learn $\boldsymbol{W}$ are sampled at 16 kHz. STFT is computed using a Hamming window that is 64 ms long with a 32 ms overlap. The parameters of the proposed scheme are set to $\lambda_1 = 2$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.8$, respectively, to attain reasonable performance. The initial estimates $\boldsymbol{Z}_j^{(0)}$, $j = 1,2,3$, are set to random matrices, and the algorithm is run for 600 iterations. The parameter of BND is set to $\beta = 0.5$ to attain reasonable performance. The initial estimate for BND is set to a matrix with all entries equal to one, and the algorithm is run for 600 iterations at each frame. As post-processing, for both system, all the entries of $\boldsymbol{H}$ are divided by the maximum value of each estimated $\boldsymbol{H}$, and the threshold which is set manually to 0.05 is used for evaluation; each entry of $\boldsymbol{H}$ which is greater (or smaller) than the threshold is considered to be active (or inactive).

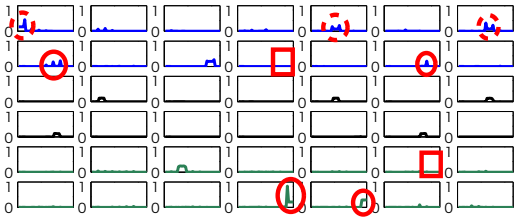
Fig. 1(a) illustrates the sound volumes of $\boldsymbol{Y}$ over 1.5 seconds approximately, and Figs. 1(b) and 1(c) depict the activation vectors of $\boldsymbol{H}$ obtained by the proposed and BND methods, respectively. In Fig. 1(c), notable incorrect estimates of activation vectors are marked in red as follows. Substitution and missed errors are indicated by squares, false alarms remaining after post-processing are indicated by solid circles, and the frames that have no errors in terms of music transcription but have inaccurate variations of sound volumes are indicated by dashed circles. The resulting $\boldsymbol{H}$ obtained by the BND method contains some components having large values (see the activation vector at the center of the bottom row in Fig. 1(c)), and



(a) Ground-truth reference



(b) The proposed scheme



(c) BND [11]

Fig. 1. Ground-truth reference and the resulting $\boldsymbol{H}$s obtained by the proposed and BND methods. The horizontal and vertical axes correspond to time and amplitudes, respectively. The pair of rows on the top (blue), in the middle (black), and at the bottom (light green) represent 14 activations for sine, triangle, and square waves, respectively.

therefore the other components representing correct activations are small due to normalization.

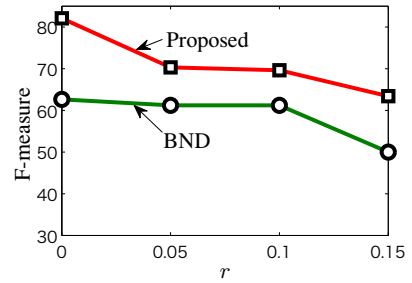
Table 1 summarizes the results in the standard evaluation metrics from the MIREX [19]. It is seen that the proposed scheme attains higher scores in $\mathcal{F}$ and $\mathcal{A}$ and lower scores in the other metrics (measuring different types of errors), meaning that it outperforms BND [11] in all the metrics. It is also seen from Fig. 1(b) that the proposed scheme selects the relevant basis vectors correctly and estimates the activation vectors accurately. This is thanks to the three regularizers, among which the temporal-continuity and the block ℓ_1 norm are not exploited in the BND method.

3.2. Case of Imperfect $\boldsymbol{W}$

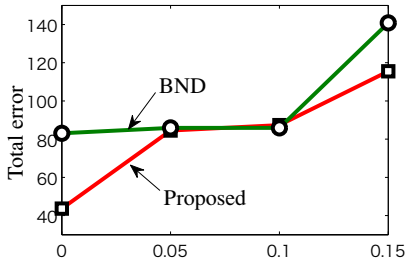
To assess the generalization ability, we conduct another simulation, considering imperfection in $\boldsymbol{W}$. To be specific, each component $w_{m,l}$ of $\boldsymbol{W}$ is corrupted by its proportional amount of noise $u_r w_{m,l}$ where $u_r \sim \mathcal{U}(-r, r)$ for $r = 0, 0.05, 0.10, 0.15$. Here, $\mathcal{U}(-r, r)$ denotes the uniform distribution in the interval $(-r, r)$. All the other conditions, including the parameters, the number of iterations, and the threshold, are the same as in Section 3.1. Fig. 2 plots the results in F-measure and the total error [19] against r. It is seen that the proposed scheme outperforms the BND method also in generalization ability.

Table 1. Transcription evaluation from the MIREX [19]. Each metric stands for F-measure, Accuracy, substitutions, misses, false alarms, and total errors.

Algorithm	$\mathcal{F}$	$\mathcal{A}$	$\mathcal{E}_{\text{subs}}$	$\mathcal{E}_{\text{miss}}$	$\mathcal{E}_{\text{fals}}$	$\mathcal{E}_{\text{tot}}$
Proposed	**82.1**	**69.6**	**0**	**0**	**43.7**	**43.7**
BND [11]	62.6	45.6	12.7	7.0	63.4	83.1



(a)



(b)

Fig. 2. Comparisons in (a) F-measure and (b) total error for imperfect $\boldsymbol{W}$.

4. CONCLUDING REMARKS

This paper presented a simple and flexible approach to supervised NMF based on a sparse optimization problem in the context of music transcription and musical instrument recognition. We designed the cost function by penalizing the data-fidelity term with four convex functions three of which are nonsmooth: (i) the nonnegativity-enforcing indicator function, (ii) the block ℓ_1 norm, (iii) the temporal-continuity term, and (iv) the ℓ_1 norm. The GFBS algorithm ensures global convergence to a minimizer of the cost function. The simulation results showed that block ℓ_1 norm is effective in selecting the relevant basis vectors from the basis matrix and that the temporal-continuity term and the ℓ_1 norm are effective in estimating the activation vectors correctly. It should be remarked again that the challenging task of determining the exact number of relevant basis vectors prior to decomposition is unnecessary in the proposed scheme. It will be an interesting future work to extend the proposed scheme with divergence such as the Itakura-Saito divergence.

Acknowledgment: This work was partially supported by JSPS Grants-in-Aid (24760292) and the Support Center for Advanced Telecommunications Technology Research (SCAT) Grants-in-Aid.

5. REFERENCES

[1] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," in *Environmetrics*, 1994.

[2] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[3] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, 2000, pp. 556–562.

[4] D. Guillamet and J. Vitriá, "Analyzing non-negative matrix factorization for image classification," in *Proc. IEEE International Conference on Pattern Recognition*, 2002, pp. 116–119.

[5] P. Smaragdis, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[6] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[7] V. Y. F. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization," in *Proc. Workshop on Signal Processing with Adaptive Sparse Structured Representations*, 2009.

[8] M. D. Hoffman, D. M. Blei, and P. R. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *Proc. ICML*, 2010, pp. 641–648.

[9] H. Raguet, J. Fadili, and G. Peyré, "Generalized forward-backward splitting," *Arxiv preprint arXiv:1108.4404v3 [math.OC]*, 2011.

[10] G. Grindlay and D. P. W. Ellis, "Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1159–1169, Oct. 2011.

[11] A. Dessein, A. Cont, and G. Lemaitr, "Real-time detection of overlapping sound events with non-negative matrix factorization," in *Matrix Information Geometry*, pp. 341–371. Springer, 2012.

[12] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer, New York, 2010.

[13] H. H. Bauschke and P. L. Combettes, *Convex Analysis And Monotone Operator Theory in Hilbert Spaces*, Springer, New York: NY, 1st edition, 2011.

[14] I. Yamada, M. Yukawa, and M. Yamagishi, "Minimizing the Moreau envelope of nonsmooth convex functions over the fixed point set of certain quasi-nonexpansive mappings," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 345–390. Springer, 2011.

[15] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari, *Nonnegative Matrix and Tensor Factorizations : Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley, 1st edition, 2009.

[16] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York: NY, 19th edition, 1985.

[17] M. Goto, "Development of the RWC music database," in *Proc. ICA*, 2004, pp. 553–556.

[18] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1643–1654, Aug. 2010.

[19] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of multiple-F0 estimation and tracking systems," in *Proc. ISMIR*, 2009, pp. 315–320.