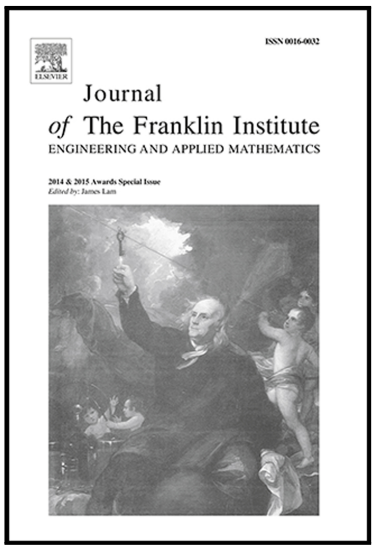


Supervised Nonnegative Matrix Factorization via Minimization of Regularized Moreau-Envelope of Divergence Function with Application to Music Transcription

Masahiro Yukawa, Hideaki Kagami

PII: S0016-0032(17)30633-6
DOI: 10.1016/j.jfranklin.2017.12.002
Reference: FI 3253

To appear in: *Journal of the Franklin Institute*

Received date: 1 November 2016
Revised date: 27 September 2017
Accepted date: 2 December 2017

Please cite this article as: Masahiro Yukawa, Hideaki Kagami, Supervised Nonnegative Matrix Factorization via Minimization of Regularized Moreau-Envelope of Divergence Function with Application to Music Transcription, *Journal of the Franklin Institute* (2017), doi: 10.1016/j.jfranklin.2017.12.002

Supervised Nonnegative Matrix Factorization via Minimization of Regularized Moreau-Envelope of Divergence Function with Application to Music Transcription

Masahiro Yukawa* and Hideaki Kagami

Department of Electronics and Electrical Engineering, Keio University, JAPAN
Hiyoshi 3-14-1, Kohoku-ku, Yokohama, Kanagawa, 223-8522 JAPAN
yukawa@elec.keio.ac.jp
** Corresponding author*

Abstract

We propose a convex-analytic approach to supervised nonnegative matrix factorization (NMF), using the Moreau envelope, a smooth approximation, of the β-divergence as a loss function. The supervised NMF problem is cast as minimization of the loss function penalized by four terms: (i) a time-continuity enhancing regularizer, (ii) the indicator function enforcing the nonnegativity, (iii) a basis-vector selector (a block ℓ_1 norm), and (iv) a sparsity-promoting regularizer. We derive a closed-form expression of the proximity operator of the sum of the three non-differentiable penalty terms (ii)–(iv). The optimization problem can thus be solved numerically by the proximal forward-backward splitting method, which requires no auxiliary variable and is therefore free from extra errors. The source number is automatically attained as an outcome of optimization. The simulation results show the efficacy of the proposed method in an application to polyphonic music transcription.

Keywords:
Nonnegative matrix factorization, supervised learning, convex analysis, proximity operator, β-divergence

1. Introduction

Face images and texts can be separated into parts and semantic features, respectively, via matrix factorization, as demonstrated experimentally by Lee and Seung [1]. Since many objects existing in nature are measured as nonnegative-valued signals, the matrix factorization problem studied there was to factorize a nonnegative matrix $\boldsymbol{Y}$ into the product of two nonnegative matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ such that $\boldsymbol{Y} \approx \boldsymbol{W}\boldsymbol{H}$. This was carried out by multiplicative updates based on the principle known widely as *majorization minimization (MM)* [2]. The problem is often called *nonnegative matrix factorization (NMF)*. Early studies related to NMF can be found, for instance, in [3, 4, 5, 6, 7, 8]. The terminology, *nonnegative matrix "approximations"*, is sometimes used to emphasize the inexactness of the factorization process [9].

Triggered by the seminal work of Lee and Seung, a significant amount of effort has been dedicated to understanding the NMF problem and to developing efficient algorithms for various applications including audio source separation, information retrieval, and clustering [10, 11, 12, 13, 14]. Several conditions are known for the uniqueness of exact factorization (excluding the scale and permutation ambiguities) [15, 16, 17, 18]. Although the visibility of the underlying structure has been enhanced, developing an efficient NMF algorithm still remains a challenging task, because the NMF is shown to be an NP-hard problem [19]. It has been reported that the gradient descent method offers better convergence behaviors, as well as additional flexibilities, than the multiplicative updates [20, 21, 22], while many of the works on NMF employ multiplicative updates by following the way of Lee and Seung. Despite the developments of powerful convex optimization tools such as *proximity operator*, those tools have only been applied to NMF in a limited number (compared to the total number of works on NMF) of works [23, 24, 25, 26, 27, 28, 29]. The discrepancy between the given nonnegative matrix (object) and the product of two variable matrices (two factors) is measured in general by some divergence function (including the squared Euclidean distance as its particular case). At the heart of the difficulty of NMF lies the nonconvexity of those "fitting" functions with respect to the two factors simultaneously.

To sidestep the difficulty of the NMF problem, we assume in the present study that one of the variable matrices is known *a priori*. More specifically, we assume the availability of a redundant dictionary. This is a practical as-

sumption in many applications including music transcription. We consider three measures: the squared Euclidean distance, the Kullback-Leibler (KL) divergence, and the dual Itakura-Saito (Dual-IS) divergence, a *dual* counterpart of the standard IS divergence. Here, the KL and Dual-IS divergences are adopted because both of them are convex and *proximable*, where a function is said to be proximable if its proximity operator can easily be computed. In the present case, the proximity operator is given in a closed form. (The standard IS divergence function is nonconvex, and hence global optimization is difficult.) Based on the three measures, we consider the following three loss functions: the squared Euclidean distance, the Moreau envelope of the KL divergence, and the Moreau envelope of the Dual-IS divergence. Here, the Moreau envelope is a smooth approximation of a possibly discontinuous convex function, and it has a close link to the proximity operator (see Section 2). (A function is said to be *smooth* if it has a Lipschitz-continuous gradient.) Indeed, *smoothing* has been studied and shown to facilitate learning [30, 31, 32, 33]. The NMF problem is cast as minimizing the loss function penalized by four terms: (i) a time-continuity encouraging regularizer, (ii) the indicator function enforcing the nonnegativity, (iii) a basis-vector selector (a block ℓ_1 norm), and (iv) a sparsity encouraging regularizer. The first penalty term is smooth, whereas all the other penalty terms are non-differentiable. We show that the three non-differentiable penalty terms are *jointly proximable*; i.e., the proximity operator of their sum can be computed easily. We remark here that the proximity operator of a sum of multiple functions is generally hard to compute, even if each function is proximable. The optimization problem can thus be solved by the proximal forward-backward splitting (PFBS) algorithm [34, 35]. In our previous studies [25, 29], the generalized forward-backward splitting algorithm [36] and the alternating direction method of multipliers (ADMM) [37, 38] have been used, both of which require auxiliary variables. In contract, the method presented in the present study requires no auxiliary variable. This avoids the extra errors caused by the introduction of the auxiliary variables, and it also leads to memory efficiency. Because of this, the proposed method achieves better performance than our previous ADMM-based method (as shown in Section 4) despite the approximation by the Moreau envelope. In addition, the PFBS algorithm is rather intuitive, since it is a generalization of the projected gradient method (PGM) [39] (see Remark 3). A remarkable advantage over the multiplicative-update approaches is that the global convergence of the iterates (not of the objective values) is rigorously guaranteed.

One of the major advantages of the supervised approach is that there is no need to determine the *source number* prior to decomposition, in addition that it is endowed with *convexity*. In the unsupervised NMF approaches, the exact number of basis vectors needs to be known prior to decomposition [40, 41]. A larger number of columns in $\boldsymbol{W}$ (than necessary) would result in producing undesirable basis vectors, while a smaller number of columns would fail in capturing desired basis vectors. The source-number determination is typically carried out by the trial and error approach [42]. In our supervised approach, the source number is determined automatically as an outcome of optimization. The simulation results show the efficacy of the proposed method in an application to polyphonic music transcription.

2. Convex Optimization Tools

Let $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ be a real Hilbert space equipped with inner product $\langle\cdot,\cdot\rangle$. The induced norm is denoted by $\|\cdot\|$. A set $S \subset \mathcal{H}$ is said to be open if for any point $x \in S$ there exists $\epsilon > 0$ such that $\{y \in S \mid \|x - y\| < \epsilon\} \subset S$. A set $S \subset \mathcal{H}$ is said to be closed if its complement set is open. A set $S \subset \mathcal{H}$ is said to be convex if $\alpha x + (1-\alpha)y \in S$ for all $(x, y, \alpha) \in \mathcal{H} \times \mathcal{H} \times [0, 1]$.

A function $f : \mathcal{H} \to (-\infty, \infty] := \mathbb{R} \cup \{\infty\}$ is said to be convex on $\mathcal{H}$ if $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$ for all $(x, y, \alpha) \in \mathsf{dom}\ f \times \mathsf{dom}\ f \times [0, 1]$, where $\mathsf{dom}\ f := \{x \in \mathcal{H} \mid f(x) < \infty\}$. If in addition $\mathsf{dom}\ f \neq \emptyset$, f is said to be a *proper convex* function. If the inequality of convex function holds with strict inequality whenever $x \neq y$, f is said to be strictly convex. A function $f : \mathcal{H} \to (-\infty, \infty]$ is said to be *lower semicontinuous* on $\mathcal{H}$ if the level set $\mathsf{lev}_{\leq a} f := \{x \in \mathcal{H} : f(x) \leq a\}$ is closed for any $a \in \mathbb{R}$. The set of proper lower-semicontinuous convex functions $f : \mathcal{H} \to (-\infty, \infty]$ is denoted by $\Gamma_0(\mathcal{H})$.

Given a nonempty closed convex set $C \subset \mathcal{H}$, define the indicator function

$$i_C(x) := \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C. \end{cases} \tag{1}$$

The function i_C is clearly discontinuous at the boundary of C, but it is lower semicontinuous because $\mathsf{lev}_{\leq a} i_C = C$ if $a \geq 0$; $\mathsf{lev}_{\leq a} i_C = \emptyset$ if $a < 0$. If a function is continuous, it is lower semicontinuous in general.

We denote by $I : \mathcal{H} \to \mathcal{H}$ the identity operator which maps any vector $\boldsymbol{x} \in \mathcal{H}$ to the $\boldsymbol{x}$ itself. Given a pair of mappings $T_1 : \mathcal{H} \to \mathcal{H}$ and $T_2 : \mathcal{H} \to \mathcal{H}$, their composition is denoted by $T_2 \circ T_1$. A mapping $T : \mathcal{H} \to \mathcal{H}$ is said to

be Lipschitz continuous with constant $\eta > 0$ (or η-Lipschitz for short) if for any $x, y \in \mathcal{H}$

$$\|T(x) - T(y)\| \leq \eta \|x - y\|. \tag{2}$$

A 1-Lipschitz mapping is said to be nonexpansive. A point that is "fixed" under the operation of $T : \mathcal{H} \to \mathcal{H}$ (i.e. a point $x \in \mathcal{H}$ such that $T(x) = x$) is called a fixed point of T. The set of fixed points is denoted as $\mathsf{Fix}(T) := \{x \in \mathcal{H} \mid T(x) = x\}$.

A matrix (a Euclidean vector) is denoted by an upper-case (lower-case) bold letter such as $\boldsymbol{A}$ ($\boldsymbol{a}$). The (i, j) entry (the ith entry) of a matrix $\boldsymbol{A}$ (a Euclidean vector $\boldsymbol{a}$) is denoted as $a_{i,j}$ (a_i). We denote by $[\cdot]_i$ the ith entry of a vector, and by $[\cdot]_{i:j}$ ($i \leq j$) the length-$(j - i + 1)$ subvector accommodating the ith to jth entries. We denote by $(\cdot)^\mathsf{T}$ the *transpose* of a matrix/vector, and by $\sigma_{\max}(\cdot)$ the largest singular value of a matrix. We denote by $\mathbf{1}$ and $\mathbf{1}_{n\times m}$ the Euclidean vector and the $n \times m$ matrix of ones, respectively, by $\boldsymbol{I}_n$ the $n \times n$ identity matrix, and by $\mathbf{0}$ and $\boldsymbol{O}_{n\times m}$ the zero Euclidean vector and the $n \times m$ zero matrix, respectively. We denote by $\|\cdot\|_1$ and $\|\cdot\|_2$ the ℓ_1 and ℓ_2 norms of a Euclidean vector, respectively. We denote by $\|\cdot\|_\mathsf{F}$ the Frobenius norm of a matrix.

In the following, we introduce the divergence function, and then present two notions of convex optimization tools: *proximity operator* and *Moreau envelope*, which are closely related to each other [43, 44].

2.1. Divergence functions

Given a set S, a nonnegative function $d : S \times S \mapsto [0, \infty)$ is called *divergence* if $d(x, y) = 0 \Leftrightarrow x = y$. Typically, a continuous function is used as a divergence. In this case, a divergence function can be used as a measure of "fitness" (or the discrepancy between two points). Note however that divergence may violate two of the axioms of metric distance: the symmetry and the triangular inequality. A divergence function therefore gives a directional distance in general, and thus the notation $d(x|y)$, $d(x||y)$, or $d(x; y)$ is often preferred rather than $d(x, y)$ to avoid confusion. We adopt the notation $d(x||y)$ in this paper.

Given a strictly convex function $\phi \in \Gamma_0(\mathcal{H})$ that is differentiable (Gâteaux differentiable in general [43]) over the interior of $\mathsf{dom}\phi$, the Bregman diver-

gence[1] is defined as follows [50]:

$$d_\phi(y||x) := \phi(y) - \phi(x) - \langle \nabla\phi(x), y - x\rangle, \quad x, y \in \mathcal{H}. \tag{3}$$

The Bregman divergence includes, as its special cases, the α-divergence [51, 52], β-divergence [53, 54, 55], the KL divergence, the logistic loss, Mahalanobis distance, etc.

In the case that $\mathcal{H} := \mathbb{R}$, letting

$$\phi(x) := \phi_\beta(x) := \begin{cases} -\log x + x - 1 & \text{if } \beta = 0 \\ x \log x - x + 1 & \text{if } \beta = 1 \\ \frac{1}{\beta(\beta-1)}(x^\beta - \beta x + \beta - 1) & \text{otherwise} \end{cases} \tag{4}$$

for $x > 0$ yields the so-called β-divergence [56, 57, 58, 59]. Note here that the affine terms in (4) can be eliminated to obtain the same divergence, while those terms are included in the literature to ensure the continuity of ϕ_β in terms of β [58, 59]. The Bregman divergence $d_\phi(y||x)$ is convex with respect to the first argument y, but can be nonconvex with respect to the second argument x. Indeed, the β-divergence is convex with respect to x only for $\beta \in [1, 2]$. For $\beta \in \mathbb{R} \setminus \{0, 1\}$, the β-divergence can be written as $d_{\phi_\beta}(y||x) = \frac{1}{\beta(\beta-1)}((1-\beta)x^\beta + \beta y^\beta + \beta x^{\beta-1} y)$. For $\beta = 2$, it reduces to the squared Euclidean distance $d_{\phi_2}(y||x) = \frac{1}{2}(x-y)^2$. For $\beta = 0, 1$, the β-divergence reduces to the IS divergence (also called Burg cross entropy [58]) $d_{\phi_0}(y||x) = \frac{y}{x} - \log\left(\frac{y}{x}\right) - 1$ and the KL divergence (also called relative entropy, information divergence, etc.) $d_{\phi_1}(y||x) = y \log\left(\frac{y}{x}\right) - y + x$, of which the seed function ϕ_1 is the negative Shannon entropy.

Typically, x is regarded as a variable for a given positive constant $y > 0$. The function

$$d^{\mathsf{dual}}_{\phi_\beta}(y||x) := d_{\phi_\beta}(x||y) \tag{5}$$

is referred to as the "dual" β-divergence, while $d_{\phi_\beta}(y||x)$ is referred to particularly as the "standard" β-divergence. In particular, $d^{\mathsf{dual}}_{\phi_0}(y||x)$ is called

[1]The Bregman divergence $d_\phi(y||x)$ is asymmetric when ϕ is non-quadratic [45]. The Bregman divergence has been studied under several different conditions on ϕ such as the so-called *Legendreness* to ensure certain desirable properties [45, 46]. Suppose that $x \in \mathcal{H}$ is an interior point of $\mathsf{dom} f$. Then, $\lim_{n\to\infty} d_\phi(y_n||x) = 0$, $(y_n)_{n\in\mathbb{N}} \subset \mathsf{dom} f$, implies $\lim_{n\to\infty} \|y_n - x\| = 0$ if and only if f is *totally convex* at x; i.e., $\inf\{d_\phi(y||x) \mid y \in \mathsf{dom} f,\ \|y - x\| = t\} > 0$ for any $t > 0$ [47, 48, 49].

the Dual-IS divergence. The standard/dual β-divergence has the following remarkable property [60]: $d_{\phi_\beta}(\alpha y||\alpha x) = \alpha^\beta d_{\phi_\beta}(y||x)$ and $d^{\mathsf{dual}}_{\phi_\beta}(\alpha y||\alpha x) = \alpha^\beta d^{\mathsf{dual}}_{\phi_\beta}(y||x)$ for any $\alpha > 0$, $x > 0$, $y > 0$. Among the β-divergences, only the standard/dual IS divergence has the *shift-invariance* property; i.e., $d_{\phi_0}(\alpha y||\alpha x) = d_{\phi_0}(y||x)$ and $d^{\mathsf{dual}}_{\phi_0}(\alpha y||\alpha x) = d^{\mathsf{dual}}_{\phi_0}(y||x)$ for any $\alpha > 0$, $x > 0$, $y > 0$. In this paper, we solely consider the KL divergence $d_{\phi_1}(y||x)$ and the Dual-IS divergence $d^{\mathsf{dual}}_{\phi_0}(y||x)$, both of which are convex and admit closed-form expressions for the proximity operators. Although a more general family of divergences called Alpha-Beta divergences has been formulated and applied to NMF [61], we do not consider it in the present study.

2.2. Proximity operator and its properties

Definition 1 (Proximity operator [62, 63, 35]). *Given any $f \in \Gamma_0(\mathcal{H})$ (see the paragraph above (1) for the definition of $\Gamma_0(\mathcal{H})$), the proximity operator of f of index $\gamma > 0$ is defined as*

$$\mathsf{prox}_{\gamma f}(x) := \underset{y\in\mathcal{H}}{\operatorname{argmin}} \left(f(y) + \frac{1}{2\gamma}\|x-y\|^2 \right), \quad x \in \mathcal{H}.$$

The uniqueness and existence of the minimizer is ensured due to the facts, respectively, that (i) $\|\cdot\|^2$ is strictly convex, and (ii) $\|\boldsymbol{x}\|^2 \to +\infty$ as $\|\boldsymbol{x}\| \to +\infty$ (coercivity of $\|\cdot\|^2$).

Definition 2 (Subgradient, subdifferential [43]). *Given $x \in \mathcal{H}$ and $f \in \Gamma_0(\mathcal{H})$, a vector $z \in \mathcal{H}$ satisfying*

$$\langle y - x, z\rangle + f(x) \leq f(y), \quad \forall y \in \mathcal{H}, \tag{6}$$

is called a subgradient of f at x. The set

$$\partial f(x) := \{z \in \mathcal{H} \mid z \text{ satisfies } (6)\} \tag{7}$$

of subgradients is called the subdifferential of f at x. If f is continuous, it is ensured that $\partial f(x) \neq \emptyset$. Subgradient is a general notion of the gradient for nondifferentiable function; $\partial f(x) = \{\nabla f(x)\}$ if f is (Gâteaux) differentiable with (Gâteaux) derivative ∇f [43].

Some useful properties of proximity operator are shown below (see also Remark 2).

Fact 1 (Selected properties of proximity operator [62, 63, 35]). *Let $f \in \Gamma_0(\mathcal{H})$ and any $\gamma > 0$. Then, the following statements hold.*

1. *(Fixed point set coincides with the set of minimizers)*

$$\mathsf{Fix}(\mathsf{prox}_{\gamma f}) = \operatorname*{argmin}_{x \in \mathcal{H}} f(x). \tag{8}$$

2. *(Connection to subdifferential — proximity operator is the resolvent of the subdifferential)*

$$\mathsf{prox}_{\gamma f} = (I + \gamma \partial f)^{-1}, \tag{9}$$

or equivalently,

$$\frac{1}{\gamma}(x - \mathsf{prox}_{\gamma f}(x)) \in \partial f(\mathsf{prox}_{\gamma f}(x)), \quad x \in \mathcal{H}. \tag{10}$$

3. *(Firm nonexpansivity — a generalization of metric projection) For any $x, y \in \mathcal{H}$,*

$$\left\|\mathsf{prox}_{\gamma f}(x) - \mathsf{prox}_{\gamma f}(y)\right\|^2 + \left\|(I - \mathsf{prox}_{\gamma f})(x) - (I - \mathsf{prox}_{\gamma f})(y)\right\|^2 \leq \|x - y\|^2. \tag{11}$$

In particular, if y is a minimizer of f, Fact 1.1 implies $\mathsf{prox}_{\gamma f}(y) = y$, and hence

$$\|x - y\|^2 - \left\|\mathsf{prox}_{\gamma f}(x) - y\right\|^2 \geq \left\|x - \mathsf{prox}_{\gamma f}(x)\right\|^2. \tag{12}$$

Note that (11) immediately implies the nonexpansivity $\left\|\mathsf{prox}_{\gamma f}(x) - \mathsf{prox}_{\gamma f}(y)\right\| \leq \|x - y\|$.

Example 1 (Proximity operator [35, 64, 29]).

1. *(Indicator function) Given any nonempty closed convex set $C \subset \mathcal{H}$ and any $\gamma > 0$,*

$$\mathsf{prox}_{\gamma i_C}(x) = \operatorname*{argmin}_{y \in C} \|x - y\| =: P_C(x), \quad x \in \mathcal{H}.$$

If, in particular, $\mathcal{H} := \mathbb{R}^n$ and $C := \mathbb{R}^n_+$, then

$$[P_{\mathbb{R}^n_+}(\boldsymbol{x})]_i := \max\{x_i, 0\}, \quad \boldsymbol{x} \in \mathbb{R}^n, \ i = 1, 2, \cdots, n. \tag{13}$$

2. *(Norm) Given any* $\gamma > 0$,

$$\mathsf{prox}_{\gamma\|\cdot\|}(x) = \begin{cases} \left(1 - \dfrac{\gamma}{\|x\|}\right) x & \text{if } \|x\| > \gamma \\ \theta & \text{if } \|x\| \leq \gamma \end{cases} \quad x \in \mathcal{H}, \tag{14}$$

where $\theta \in \mathcal{H}$ *is the zero vector of* $\mathcal{H}$*. If, in particular,* $\mathcal{H} := \mathbb{R}$*, then it reduces to the soft-thresholding (or shrinkage) operator*

$$\mathsf{prox}_{\gamma|\cdot|}(x) = \mathsf{sign}(x)\max\{|x| - \gamma, 0\}, \quad x \in \mathbb{R}. \tag{15}$$

where $\mathsf{sign}(\cdot)$ *is the signum function that returns* 1 *if the argument is nonnegative valued, and* 0 *otherwise.*

3. *(KL divergence) Given any* $\gamma > 0$ *and any* $y > 0$*, extend the range of the variable* x *to* $\mathbb{R}$ *as follows:*

$$d_{\phi_1}(y||x) := \begin{cases} y\log\left(\frac{y}{x}\right) - y + x & \text{if } x > 0 \\ +\infty & \text{if } x \leq 0. \end{cases} \tag{16}$$

Then, $d_{\phi_1}(y||\cdot) \in \Gamma_0(\mathbb{R})$ *(the lower semicontinuity can be verified by noting that* $\lim_{x\downarrow 0} d_{\phi_1}(y||x) = +\infty$*), and*

$$\mathsf{prox}_{\gamma d_{\phi_1}(y||\cdot)}(x) = \{p > 0 \mid p^2 + (\gamma - x)p = \gamma y\}, \quad x \in \mathbb{R}. \tag{17}$$

4. *(Dual-IS divergence) Given any* $\gamma > 0$ *and any* $y > 0$*, extend the range of the variable* x *to* $\mathbb{R}$ *as follows:*

$$d^{\mathsf{dual}}_{\phi_0}(y||x) := \begin{cases} d_{\phi_0}(x||y) = \frac{x}{y} - \log\left(\frac{x}{y}\right) - 1 & \text{if } x > 0 \\ +\infty & \text{if } x \leq 0. \end{cases} \tag{18}$$

Then, $d^{\mathsf{dual}}_{\phi_0}(y||\cdot) \in \Gamma_0(\mathbb{R})$ *(the lower semicontinuity can be verified by noting that* $\lim_{x\downarrow 0} d^{\mathsf{dual}}_{\phi_0}(y||x) = +\infty$*), and*

$$\mathsf{prox}_{\gamma d^{\mathsf{dual}}_{\phi_0}(y||\cdot)}(x) = \{p > 0 \mid p^2 + (\gamma y^{-1} - x)p = \gamma\}, \quad x \in \mathbb{R}. \tag{19}$$

The proximity operator is a generalization of the metric projection, as can be seen from Example 1.1. The following lemma is useful to compute the proximity operator of a separable function.

Lemma 1 (Proximity operator of separable function in product space [35]). *Let* $(\mathcal{H}^m, \langle\cdot,\cdot\rangle_{\mathcal{H}^m})$ *be the product space equipped with inner product* $\langle x, y\rangle_{\mathcal{H}^m} := \sum_{i=1}^m \langle x_i, y_i\rangle$ *for* $x = (x_i)_{i=1}^m \in \mathcal{H}^m$*,* $y = (y_i)_{i=1}^m \in \mathcal{H}^m$*. Let* $f(x) := \sum_{i=1}^m f_i(x_i)$ *for* $x = (x_i)_{i=1}^m$*, where* $f_i \in \Gamma_0(\mathcal{H})$*,* $i = 1, 2, \cdots, m$*. Then,* $\mathsf{prox}_f(x) = (\mathsf{prox}_{f_i}(x_i))_{i=1}^m$*.*

2.3. Moreau envelope and its properties

Definition 3 (Moreau envelope [62, 63, 35]). *Given a function $f \in \Gamma_0(\mathcal{H})$, its Moreau envelope of index $\gamma \in (0, \infty)$ is defined as follows:*

$$ {}^{\gamma}f : \mathcal{H} \to \mathbb{R},\ x \mapsto \min_{y \in \mathcal{H}} \left(f(y) + \frac{1}{2\gamma} \|x - y\|^2 \right). \tag{20}$$

Moreau envelope provides a smooth approximation of a potentially discontinuous convex function with nice properties as shown below.

Fact 2 (Selected properties of Moreau envelope [62, 63, 35, 44]).

1. *(Lower boundedness)*

$$ {}^{\gamma}f(x) \leq f(x), \quad \forall x \in \mathcal{H}. \tag{21}$$

2. *(Convergence) The pointwise convergence holds in general:*

$$ \lim_{\gamma \downarrow 0} {}^{\gamma}f(x) = f(x), \quad \forall x \in \mathsf{dom} f. \tag{22}$$

   If, in particular, f is uniformly continuous on a bounded subset $S \subset \mathsf{dom} f$, the convergence is uniform:

$$ \lim_{\gamma \downarrow 0} \sup_{x \in S} |f(x) - {}^{\gamma}f(x)| = 0. \tag{23}$$

3. *(Convexity and differentiability) The function ${}^{\gamma}f$ is convex and differentiable.*
4. *(Lipschitz-continuous gradient) The gradient is given by*

$$ \nabla {}^{\gamma}f(x) = \frac{1}{\gamma} \left(x - \mathsf{prox}_{\gamma f}(x) \right), \tag{24}$$

   which is $\frac{1}{\gamma}$-Lipschitz.
5. *(Preservation of global minimizers)*

$$ \operatorname*{argmin}_{x \in \mathcal{H}} {}^{\gamma}f(x) = \operatorname*{argmin}_{x \in \mathcal{H}} f(x) = \mathsf{Fix}(\mathsf{prox}_{\gamma f}). \tag{25}$$

3. Proposed NMF Algorithm

Our problem formulation involves a divergence-based loss function and multiple penalty terms. As a loss function, we consider the KL and Dual-IS divergences as well as the squared Euclidean distance. We repeat here that the KL and Dual-IS divergences are convex and proximable. The penalty terms employed include a continuity-enhancing regularizer, an indicator function that enforces the nonnegativity, a block ℓ_1-norm regularizer for basis-vector selection, and a sparsity-promoting regularizer. The simultaneous use of all the above regularizers yields significant improvements in performance, as shown by simulations in Section 4. The entire objective function could be optimized by the standard solver of ADMM, as studied in [29]. ADMM, however, employs auxiliary variables, and the errors in those auxiliary variables typically cause some extra errors in the primal variables. In addition, the use of auxiliary variables increases the computational complexity and memory requirements. We therefore reformulate the problem by smoothing the divergence-based loss functions by means of Moreau envelope, which has been introduced in Section 2.3. To make the simple auxiliary-variable-free solver of PFBS applicable to the reformulated problem, we show that the nonsmooth penalty functions (the last three penalty terms that are raised above) are jointly proximable.

3.1. Problem formulation and optimization algorithm

Let $\boldsymbol{Y} \in \mathbb{R}_+^{M\times N}$ be a given $M \times N$ nonnegative matrix to be factorized, where $\mathbb{R}_+$ denotes the set of nonnegative real numbers. We assume the availability of $\boldsymbol{W} \in \mathbb{R}^{M\times L}$ of which the column vectors form a redundant dictionary. The task is to find an $L \times N$ matrix $\boldsymbol{H} \in \mathbb{R}_+^{L\times N}$ such that $\boldsymbol{Y} \approx \boldsymbol{W}\boldsymbol{H}$. The discrepancy between $\boldsymbol{Y}$ and $\boldsymbol{W}\boldsymbol{H}$ is measured by either the squared Euclidean distance or the KL/Dual-IS divergence introduced in Section 2.1.

We consider the Hilbert space $\mathcal{H} := \mathbb{R}^{L\times N}$ equipped with the inner product $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle := \sum_{l=1}^{L}\sum_{n=1}^{N} x_{l,n}y_{l,n} = \mathsf{trace}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{Y})$, $\boldsymbol{X}, \boldsymbol{Y} \in \mathcal{H}$. The induced norm $\|\boldsymbol{X}\| := \sqrt{\langle \boldsymbol{X}, \boldsymbol{X}\rangle}$, $\boldsymbol{X} \in \mathcal{H}$, is the Frobenius norm. Let $\boldsymbol{h}_l \in \mathbb{R}^N$, $l = 1, 2, \cdots, L$, be the lth column of $\boldsymbol{H}^{\mathsf{T}}$; i.e., $\boldsymbol{H}^{\mathsf{T}} = [\boldsymbol{h}_1\ \boldsymbol{h}_2\ \cdots\ \boldsymbol{h}_L]$. We denote by $C := \mathbb{R}_+^{L\times N} \subset \mathcal{H}$ the constraint set imposing the nonnegativity.

We formulate the supervised NMF problem as the following convex opti-

mization problem:

$$\min_{\boldsymbol{H}\in\mathcal{H}} \quad \underbrace{\ell(\boldsymbol{Y},\boldsymbol{W}\boldsymbol{H}) + \frac{\lambda_1}{2}\sum_{l=1}^{L}\sum_{n=1}^{N-1}(h_{l,n+1}-h_{l,n})^2}_{=:\,\varphi(\boldsymbol{H})\quad\text{(smooth)}} + \underbrace{i_C(\boldsymbol{H}) + \lambda_2\sum_{l=1}^{L}\|\boldsymbol{h}_l\|_2 + \lambda_3\sum_{l=1}^{L}\sum_{n=1}^{N}|h_{l,n}|}_{=:\,\psi(\boldsymbol{H})\quad\text{(jointly proximable)}}. \tag{26}$$

Each term of (26) is detailed below.

(a) **Loss function:** We consider the squared Euclidean distance, the KL divergence, and the Dual-IS divergence. Since the loss function is nonsmooth (the gradient is not Lipschitz continuous) for the KL and Dual-IS divergences, we use their Moreau envelopes.

(b) **Continuity-enhancing regularizer:** The term $\sum_{l=1}^{L}\sum_{n=1}^{N-1}(h_{l,n+1}-h_{l,n})^2$ enhances the continuity among the adjacent components of each row vector of $\boldsymbol{H}$. We solely consider this particular regularizer because of its convexity as well as its simplicity, although some other regularizers have also been proposed for similar purposes, such as the one that uses hidden Markov models [65]. In the context of music transcription, it actually suppresses the differences among the values in adjacent frames of each activation vector, thereby enhancing the temporal continuity [13] (see Section 4).

(c) **Nonnegativity constraint:** The term $i_C(\boldsymbol{H})$ is the indicator function enforcing the matrix $\boldsymbol{H}$ to be nonnegative.

(d) **Basis-vector selecting regularizer:** The term $\sum_{l=1}^{L}\|\boldsymbol{h}_l\|_2$ is the block ℓ_1-norm penalty for selecting the necessary columns to express $\boldsymbol{Y}$ from the redundant dictionary matrix $\boldsymbol{W}$.

(e) **Sparsity-promoting regularizer:** The term $\sum_{l=1}^{L}\sum_{n=1}^{N}|h_{l,n}|$ enhances the global sparseness of the solution. It can also be regarded as a sum of the ℓ_1 norms of the column/row vectors, implying that sparseness of each column/row vector will be encouraged. Indeed, in music transcription for instance, both column sparseness and row sparseness are desired because only a few pitches are active at a given time frame while each pitch could typically be active only at short consecutive

time-frames. This regularizer together with the temporal continuity contributes to estimating the correct activation vectors.

The problem in (26) can be solved efficiently by the PFBS algorithm [34, 35]: for an arbitrary initial matrix $\boldsymbol{H}_0 \in \mathcal{H}$, generate $(\boldsymbol{H}_k)_{k\in\mathbb{N}}$ by

$$\boldsymbol{H}_{k+1} := \mathsf{prox}_{\mu\psi}\left(\boldsymbol{H}_k - \mu\nabla\varphi(\boldsymbol{H}_k)\right), \quad k \in \mathbb{N}, \tag{27}$$

where $\mu := \eta_\varphi^{-1}\delta \in (0, 2/\eta_\varphi)$ with the Lipschitz constant $\eta_\varphi > 0$ of $\nabla\varphi$ and a constant $\delta \in (0, 2)$. Detailed discussions about the PFBS algorithm will be given in Remarks 2 and 3 in Section 3.4.

3.2. Computation of the gradient

We show how to compute the gradient of the smooth part $\varphi(\boldsymbol{H}) := \Phi(\boldsymbol{H}) + \lambda_1\Omega(\boldsymbol{H})$ in (26), where $\Phi(\boldsymbol{H}) := \ell(\boldsymbol{Y}, \boldsymbol{W}\boldsymbol{H})$ and

$$\Omega(\boldsymbol{H}) := \frac{1}{2}\sum_{l=1}^{L}\sum_{n=1}^{N-1}(h_{l,n+1} - h_{l,n})^2 = \frac{1}{2}\|\boldsymbol{H}\boldsymbol{A}\|_{\mathsf{F}}^2 \tag{28}$$

with

$$\boldsymbol{A} := \begin{bmatrix} \boldsymbol{I}_{N-1} \\ \boldsymbol{0}^{\mathsf{T}} \end{bmatrix} - \begin{bmatrix} \boldsymbol{0}^{\mathsf{T}} \\ \boldsymbol{I}_{N-1} \end{bmatrix}.$$

The gradient of $\Omega(\boldsymbol{H})$ is given by

$$\nabla\Omega(\boldsymbol{H}) = \boldsymbol{H}\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}, \tag{29}$$

which is $\sigma_{\max}(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A})$-Lipschitz. The gradient of $\Phi(\boldsymbol{H})$ depends on the choice of the loss function $\ell(\boldsymbol{Y}, \boldsymbol{W}\boldsymbol{H})$. The gradient $\nabla\varphi(\boldsymbol{H}) = \nabla\Phi(\boldsymbol{H}) + \lambda_1\nabla\Omega(\boldsymbol{H})$ is η_φ-Lipschitz with $\eta_\varphi = \eta_\Phi + \lambda_1\sigma_{\max}(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A})$, where we suppose that $\nabla\Phi$ is η_Φ-Lipschitz for some $\eta_\Phi > 0$.

3.2.1. Case of squared Euclidean distance

Suppose that one employs the squared Euclidean distance

$$\Phi_{\mathsf{EUC}}(\boldsymbol{H}) := \ell_{\mathsf{EUC}}(\boldsymbol{Y}, \boldsymbol{W}\boldsymbol{H}) := \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{H}\|_{\mathsf{F}}^2. \tag{30}$$

In this case, the gradient

$$\nabla\Phi_{\mathsf{EUC}}(\boldsymbol{H}) = \boldsymbol{W}^{\mathsf{T}}(\boldsymbol{W}\boldsymbol{H} - \boldsymbol{Y}) \tag{31}$$

is η_Φ-Lipschitz for $\eta_\Phi := \sigma_{\max}(\boldsymbol{W}^{\mathsf{T}}\boldsymbol{W})$. Therefore, $\Phi_{\mathsf{EUC}}(\boldsymbol{H})$ is smooth itself, and there is no need to use its Moreau envelope in this case.

3.2.2. Case of divergence functions

We denote by $d(y||x)$ either the KL divergence $d_{\phi_1}(y||x)$ or the Dual-IS divergence $d^{\mathsf{dual}}_{\phi_0}(y||x)$ defined in Example 1. Given matrices $\boldsymbol{X} \in \mathbb{R}^{M\times N}$ and $\boldsymbol{Y} \in \mathbb{R}_+^{M\times N}$, define the following separable divergences:

$$D(\boldsymbol{X}) := D(\boldsymbol{Y}||\boldsymbol{X}) := \sum_{m=1}^{M}\sum_{n=1}^{N} d(y_{m,n}||x_{m,n}). \tag{32}$$

Using the Moreau envelope ${}^{\gamma}D$ of D for some $\gamma > 0$, our loss function based on the divergence function is given by

$$\Phi_{\mathsf{D\text{-}MR}}(\boldsymbol{H}) := \ell(\boldsymbol{Y}, \boldsymbol{W}\boldsymbol{H}) := {}^{\gamma}D(\boldsymbol{W}\boldsymbol{H}) \tag{33}$$

with its gradient

$$\nabla\Phi_{\mathsf{D\text{-}MR}}(\boldsymbol{H}) = \boldsymbol{W}^{\mathsf{T}}\frac{\boldsymbol{W}\boldsymbol{H} - \mathsf{prox}_{\gamma D}(\boldsymbol{W}\boldsymbol{H})}{\gamma} \tag{34}$$

being η_Φ-Lipschitz for $\eta_\Phi := \sigma_{\max}(\boldsymbol{W}^{\mathsf{T}}\boldsymbol{W})/\gamma$. Here, (34) is verified by Fact 2.4 together with the chain rule. The term $\mathsf{prox}_{\gamma D}(\boldsymbol{W}\boldsymbol{H})$ can be computed by Lemma 1 and (17) or (19). More specifically, for the KL divergence, its (m,n) component is given by

$$\begin{aligned}[\mathsf{prox}_{\gamma D}(\boldsymbol{W}\boldsymbol{H})]_{m,n} &= \left\{p > 0 \mid p^2 + (\gamma - [\boldsymbol{W}\boldsymbol{H}]_{m,n})p - \gamma y_{m,n} = 0\right\} \\ &= \frac{[\boldsymbol{W}\boldsymbol{H}]_{m,n} - \gamma + \sqrt{([\boldsymbol{W}\boldsymbol{H}]_{m,n} - \gamma)^2 + 4\gamma y_{m,n}}}{2},\end{aligned} \tag{35}$$

where $[\boldsymbol{W}\boldsymbol{H}]_{m,n}$ is the (m,n) component of $\boldsymbol{W}\boldsymbol{H}$. The proximity operator for the Dual-IS divergence can analogously be obtained.

3.3. Computation of the proximity operator

We show that the nonsmooth function ψ in (26) is proximable (i.e., the nonsmooth penalty terms are jointly proximable). Consider the real Hilbert space $\mathbb{R}^N$ of length-N Euclidean vectors equipped with the ℓ_2 norm $\|\cdot\|_2$. We first prove the following lemma.

Lemma 2. *Let $f(\boldsymbol{x}) := i_{\mathbb{R}^N_+}(\boldsymbol{x}) + \lambda\|\boldsymbol{x}\|_2$ for $\lambda > 0$. Then,*

$$\mathsf{prox}_f(\boldsymbol{x}) = \mathsf{prox}_{\lambda\|\cdot\|_2} \circ P_{\mathbb{R}^N_+}(\boldsymbol{x}). \tag{36}$$

Proof. Although the claim can be verified by a more general result given in [66, Lemma 2.2], we present a primitive proof for convenience. Without loss of generality, we assume that $x_1 \geq x_2 \geq \cdots \geq x_N$ and that only the first $s(\leq N)$ components are nonnegative. By definition, we have

$$\begin{aligned}\mathsf{prox}_f(\boldsymbol{x}) &= \operatorname*{argmin}_{\boldsymbol{y}\in\mathbb{R}^N_+} \left(\lambda \|\boldsymbol{y}\|_2 + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right) \\ &= \operatorname*{argmin}_{y_i\geq 0} \left(\lambda\sqrt{\sum_{n=1}^N y_n^2} + \frac{1}{2}\sum_{n=1}^N (x_n - y_n)^2\right). \end{aligned} \tag{37}$$

It is clear that, when $s < N$, the minimum is achieved by $y_n = 0$ for $n > s$, implying that

$$[\mathsf{prox}_f(\boldsymbol{x})]_n = 0, \quad n > s. \tag{38}$$

The first s components of $\mathsf{prox}_f(\boldsymbol{x})$ are therefore given as

$$\begin{aligned}[\mathsf{prox}_f(\boldsymbol{x})]_{1:s} &= \operatorname*{argmin}_{y_i\geq 0} \left(\lambda\sqrt{\sum_{n=1}^s y_n^2} + \frac{1}{2}\sum_{n=1}^s (x_n - y_n)^2\right). \\ &= \mathsf{prox}_{\lambda\|\cdot\|_2}(\boldsymbol{x}_{1:s}) \in \mathbb{R}^s_+, \end{aligned} \tag{39}$$

where the second equality is due to the fact that the operator $\mathsf{prox}_{\lambda\|\cdot\|_2}$ keeps the sign of each component unchanged (see Example 1.2). On the other hand, the equality $P_{\mathbb{R}^N_+}(\boldsymbol{x}) = \begin{bmatrix}\boldsymbol{x}_{1:s}\\ \boldsymbol{0}\end{bmatrix}$ implies

$$\mathsf{prox}_{\lambda\|\cdot\|_2} \circ P_{\mathbb{R}^N_+}(\boldsymbol{x}) = \mathsf{prox}_{\lambda\|\cdot\|_2}\left(\begin{bmatrix}\boldsymbol{x}_{1:s}\\ \boldsymbol{0}\end{bmatrix}\right). \tag{40}$$

Suppose that $\left\|P_{\mathbb{R}^N_+}(\boldsymbol{x})\right\|_2 = \|\boldsymbol{x}_{1:s}\|_2 \leq \lambda$. In this case, by Example 1.2, it holds that $\mathsf{prox}_{\lambda\|\cdot\|_2}(\boldsymbol{x}_{1:s}) = \boldsymbol{0}$ ($\Rightarrow \mathsf{prox}_f(\boldsymbol{x}) = \boldsymbol{0}$ by (38) and (39)) and $\mathsf{prox}_{\lambda\|\cdot\|_2}\left(\begin{bmatrix}\boldsymbol{x}_{1:s}\\ \boldsymbol{0}\end{bmatrix}\right) = \boldsymbol{0}$. Suppose now that $\left\|P_{\mathbb{R}^N_+}(\boldsymbol{x})\right\|_2 = \|\boldsymbol{x}_{1:s}\|_2 > \lambda$. In this case, by Example 1.2, it holds that $\mathsf{prox}_{\lambda\|\cdot\|_2}\left(\begin{bmatrix}\boldsymbol{x}_{1:s}\\ \boldsymbol{0}\end{bmatrix}\right) = \left(1 - \dfrac{\lambda}{\|\boldsymbol{x}_{1:s}\|_2}\right)\begin{bmatrix}\boldsymbol{x}_{1:s}\\ \boldsymbol{0}\end{bmatrix} =$

$\left[\begin{array}{c}\left(1-\dfrac{\lambda}{\|\boldsymbol{x}_{1:s}\|_2}\right)\boldsymbol{x}_{1:s}\\ \mathbf{0}\end{array}\right] = \left[\begin{array}{c}\mathsf{prox}_{\lambda\|\cdot\|_2}(\boldsymbol{x}_{1:s})\\ \mathbf{0}\end{array}\right] = \mathsf{prox}_f(\boldsymbol{x})$, where the last equality is due to (38) and (39). This completes the proof. □

We can then prove the following proposition.

Proposition 1. *Let* $g(\boldsymbol{x}) := i_{\mathbb{R}_+^N}(\boldsymbol{x}) + \lambda\|\boldsymbol{x}\|_2 + \nu\|\boldsymbol{x}\|_1$ *for* $\lambda, \nu > 0$. *Then,*

$$\mathsf{prox}_g(\boldsymbol{x}) = \mathsf{prox}_{\lambda\|\cdot\|_2} \circ P_{\mathbb{R}_+^N}(\boldsymbol{x} - \nu\mathbf{1}). \tag{41}$$

Proof. By definition of the proximity operator, we obtain

$$\begin{aligned}\mathsf{prox}_g(\boldsymbol{x}) &= \underset{\boldsymbol{y}\in\mathbb{R}_+^N}{\operatorname{argmin}} \left(\lambda\|\boldsymbol{y}\|_2 + \nu\|\boldsymbol{y}\|_1 + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right)\\ &= \underset{\boldsymbol{y}\in\mathbb{R}_+^N}{\operatorname{argmin}} \left(\lambda\|\boldsymbol{y}\|_2 + \nu\mathbf{1}^\mathsf{T}\boldsymbol{y} + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right)\\ &= \underset{\boldsymbol{y}\in\mathbb{R}^N}{\operatorname{argmin}} \left(i_{\mathbb{R}_+^N}(\boldsymbol{y}) + \lambda\|\boldsymbol{y}\|_2 + \frac{1}{2}\|\boldsymbol{y}-(\boldsymbol{x}-\nu\mathbf{1})\|_2^2\right)\\ &= \mathsf{prox}_{i_{\mathbb{R}_+^N}+\lambda\|\cdot\|_2}(\boldsymbol{x}-\nu\mathbf{1}),\end{aligned} \tag{42}$$

which together with Lemma 2 verifies (41). □

Under the light of Lemma 1, the matrix $\boldsymbol{H}_{k+1} = \mathsf{prox}_{\mu\psi}(\boldsymbol{G}_k)$, where $\boldsymbol{G}_k := \boldsymbol{H}_k - \mu\nabla\varphi(\boldsymbol{H}_k)$, can be obtained by applying Proposition 1 to each column of $\boldsymbol{G}_k^\mathsf{T}$ separately. The proposed algorithm based on forward-backward splitting is summarized in Table 1.

3.4. Discussion

We mention that the proximal forward-backward splitting algorithm presented in (27) is a particular example of Mann iterates [67, 68] (see [44]). The following global convergence result is immediately obtained.

Theorem 1. *The sequence* $(\boldsymbol{H}_k)_{k\in\mathbb{N}}$ *generated by the proposed algorithm is convergent to a minimizer of the objective function* $\varphi(\boldsymbol{H})+\psi(\boldsymbol{H})$, *or equivalently, to a minimizer of* $\ell(\boldsymbol{H})+\lambda_1\Omega(\boldsymbol{H})+\lambda_2\sum_{l=1}^L\|\boldsymbol{h}_l\|_2+\lambda_3\sum_{l=1}^L\sum_{n=1}^N|h_{l,n}|$ *over the nonnegativity constraint set* C. *Here,* $\ell(\boldsymbol{H})$ *is the squared Euclidean distance, or the Moreau envelope of the KL/Dual-IS divergence.*

Remark 1.

Table 1: Proposed supervised-NMF algorithm based on forward-backward splitting.

Requirement:
$\lambda_1, \lambda_2, \lambda_3 > 0$, $\delta \in (0, 2)$
$\gamma > 0$ (the index of the Moreau envelope)

Initialization and stepsize setting:
$\boldsymbol{H}_0 \in \mathbb{R}^{L\times N}$: arbitrary
$\mu := (\eta_\Phi + \lambda_1 \sigma_{\max}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A}))^{-1}\delta$ with
$\eta_\Phi := \begin{cases} \sigma_{\max}(\boldsymbol{W}^\mathsf{T}\boldsymbol{W}) & \text{(case of the squared Euclidean distance)} \\ \sigma_{\max}(\boldsymbol{W}^\mathsf{T}\boldsymbol{W})/\gamma & \text{(case of KL/Dual-IS divergence)} \end{cases}$

Update:
- Case of the squared Euclidean distance:
  $\nabla\varphi(\boldsymbol{H}_k) := \lambda_1 \boldsymbol{H}_k\boldsymbol{A}\boldsymbol{A}^\mathsf{T} + \boldsymbol{W}^\mathsf{T}(\boldsymbol{W}\boldsymbol{H}_k - \boldsymbol{Y})$
- Case of KL/Dual-IS divergence:
  $\nabla\varphi(\boldsymbol{H}_k) := \lambda_1 \boldsymbol{H}_k\boldsymbol{A}\boldsymbol{A}^\mathsf{T} + \boldsymbol{W}^\mathsf{T}\dfrac{\boldsymbol{W}\boldsymbol{H}_k - \mathsf{prox}_{\gamma D}(\boldsymbol{W}\boldsymbol{H}_k)}{\gamma}$ See (35) and (19).

$\boldsymbol{G}_k := \boldsymbol{H}_k - \mu\nabla\varphi(\boldsymbol{H}_k)$
$\boldsymbol{B}_k := [\boldsymbol{b}_{k,1}\ \boldsymbol{b}_{k,2}\ \cdots\ \boldsymbol{b}_{k,L}]^\mathsf{T} := P_{\mathbb{R}_+^{L\times N}}(\boldsymbol{G}_k - \mu\lambda_3 \mathbf{1}_{L\times N})$ See (13).
$\boldsymbol{h}_{k+1,l} := \mathsf{prox}_{\mu\lambda_2\|\cdot\|_2}(\boldsymbol{b}_{k,l}), \quad l = 1, 2, \cdots, L$ See (14).
$\boldsymbol{H}_{k+1} := [\boldsymbol{h}_{k+1,1}\ \boldsymbol{h}_{k+1,2}\ \cdots\ \boldsymbol{h}_{k+1,L}]^\mathsf{T}$

(a) *On uniqueness and existence of minimizer of $\varphi(\boldsymbol{H}) + \psi(\boldsymbol{H})$: Since $\sum_{l=1}^L \|\boldsymbol{h}_l\|_2$ and $\sum_{l=1}^L\sum_{n=1}^N |h_{l,n}|$ are coercive, the existence of a minimizer is automatically guaranteed. The uniqueness, on the other hand, depends on the choice of loss function and the dictionary $\boldsymbol{W}$ in general. Fortunately, however, the squared Euclidean distance, the KL divergence $d_{\phi_1}(y||\cdot)$, and the Dual-IS divergence $d_{\phi_0}^{\mathsf{dual}}(y||\cdot)$ (see Example 1) are all strictly convex. Hence, their associated loss functions (without the Moreau-envelope approximation) are also strictly convex if the mapping $\boldsymbol{W} : L \to M$ is injective (this is the case where $M \geq L$ and the columns of $\boldsymbol{W}$ are linearly independent). In this case, the Moreau-envelope of the KL/Dual-IS divergence has a unique minimizer which coincides with the minimizer of the divergence itself (see Fact 2.5). This does not, however, ensure the uniqueness of minimizer of $\varphi(\boldsymbol{H})+\psi(\boldsymbol{H})$ because the strict convexity of the Moreau-envelope is not ensured in*

general.

(b) *The Moore-Penrose pseudo-inverse: Since the matrices* $\boldsymbol{Y}$ *and* $\boldsymbol{W}$*, one may wonder what will happen if we use* $\boldsymbol{H} = \boldsymbol{W}^{\dagger}\boldsymbol{Y}$ *with the Moore-Penrose pseudo-inverse* $\boldsymbol{W}^{\dagger}$ *of* $\boldsymbol{W}$*. This approach, however, tends to yield unsuccessful decomposition; i.e., the resulting* $\boldsymbol{H}$ *tends to contain negative entries and to be a dense matrix as well.*

(c) *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* $\boldsymbol{W}$ *[69]. Their method relies also on an additional statistical assumption that "a suitably normalized magnitude spectrum can be modeled as a joint distribution over time and frequency". In contrast, our approach is deterministic (relying on no statistical assumption).*

Dessein et al. have proposed a real-time event-detection scheme based on convex quadratic programming [70]. The approach has been developed for real-time processing and could fail in estimating activation vectors correctly even in simple numerical simulations [25], 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 fairly flexible in the sense that other possible convex penalties can be easily incorporated. Other previous studies on supervised NMF can be found, e.g., in [71, 72, 73, 74, 75].*

(d) *On the choice of* γ *in the case of divergence-based loss functions: The parameter* γ *controls the closeness of the Moreau envelope* ${}^{\gamma}f$ *to its source function* f*. Specifically, the smaller* γ *is, the closer to* f *the Moreau envelope is. This, however, does not imply that the smaller the better regarding* γ *because a too small* γ *results in very slow convergence, as suggested by the following observation:* $\mu \downarrow 0$ *as* $\gamma \downarrow 0$ $(\Leftarrow \eta_{\Phi} = \sigma_{\max}(\boldsymbol{W}^{\mathsf{T}}\boldsymbol{W})/\gamma \to \infty$ *as* $\gamma \downarrow 0)$*. Therefore,* γ *needs to be tuned appropriately. The performance of the proposed method is insensitive indeed to the choice of* γ*, and thus its tuning is relatively easy compared to* λ_2 *and* λ_3*, as mentioned also in Section 4.*

(e) *Computational complexity: At each iteration, the proposed algorithm requires* $(MN + 4N + 3)L$ *multiplications when the squared Euclidean distance is employed, and* $(2MN + 4N + 3)L + 4MN$ *multiplications when the KL/Dual-IS divergence is employed. Hence, the total num-*

ber of multiplications for optimization is (the per-iteration cost) × (# iterations) + (initialization cost). The initialization cost is typically negligible compared to the other term because the matrix $\boldsymbol{A}$ has a simple structure and $\sigma_{\max}(\boldsymbol{W}^{\mathsf{T}}\boldsymbol{W}) = \sigma_{\max}(\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}})$ is obtained by solving an eigenvalue problem of an $r \times r$ matrix, where $r := \min\{L, N\}$. (In typical situations, $L \ll N$.)

Remark 2 (Proximity operator and Theorem 1). *Fact 1.2 implies an essential difference from the subgradient method. Specifically, the direction of the displacement vector $\mathsf{prox}_{\gamma f}(x) - x$ is given by an anti-subgradient of f at the point $\mathsf{prox}_{\gamma f}(x)$, whereas the direction of the displacement vector for the subgradient method is given by an anti-subgradient of f at x. Namely, the proximity operator leverages the local information about f at the point after operating $\mathsf{prox}_{\gamma f}$. Using Fact 1.2, we can show that [43, 44]*

$$\mathsf{Fix}(\mathsf{prox}_{\mu\psi}\left(I - \mu\nabla\varphi\right)) = \operatorname*{argmin}_{\boldsymbol{H}\in\mathcal{H}} \varphi(\boldsymbol{H}) + \psi(\boldsymbol{H}). \tag{43}$$

Equation (12) in Fact 1.3 states that $\mathsf{prox}_{\gamma f}$ pushes our estimate x towards every minimizer y of f. More than that, the difference $\|x-y\|^2 - \left\|\mathsf{prox}_{\gamma f}(x) - y\right\|^2$ is ensured to be no smaller than the (squared) travelling distance $\left\|x - \mathsf{prox}_{\gamma f}(x)\right\|^2$. This property is a particular case of the so-called strongly attracting (quasi-)nonexpansiveness. Intuitively, the larger our step is, the closer we get to every minimizer. In general, (12) is required for every $y \in \mathsf{Fix}(\mathsf{prox}_{\gamma f})$ to ensure that the iterates converge to a fixed point of the operator $\mathsf{prox}_{\gamma f}$. In the present case, Fact 1.1 ensures that every fixed point is a minimizer of f, and vice versa. The strongly attracting (quasi-)nonexpansiveness is essentially the key ingredients of Mann iterates,[2] and Theorem 1 can be verified with those arguments (see [44]).

Remark 3 (PFBS and the projected gradient method). *If ψ is an indicator function in (27), the proximity operator boils down to the projection operator (see Example 1.1), and thus the PFBS algorithm reduces to the PGM [39], a standard iterative solver for constrained optimization problems. Hence, PFBS is a generalization of PGM, and this gives an intuition about*

[2]Mann iterate is usually defined with averaged (quasi-)nonexpansive mapping. In fact, averaged (quasi-)nonexpansiveness is an equivalent notion to strongly attracting (quasi-) nonexpansiveness as long as a fixed point exists [43, 44].

PFBS. Indeed, PGM first shifts the current estimate in the steepest descent direction, and then pulls it back to the constraint set. PFBS performs exactly the same operation as its first step, and then pushes our estimate towards the set of minimizers of ψ (see Remark 2). In other words, it seeks to minimize the smooth function φ and the nonsmooth function ψ alternately by means of the steepest-descent shift and the proximity operator, respectively. We emphasize here that, after every operation of the gradient and proximity steps, our estimate is ensured to be closer to every minimizer of the objective function $\varphi + \psi$ (Fejér monotonicity).

4. Simulation Results

We first apply the proposed method to artificial data, and then apply it to music transcription problems. The proposed method is compared with the existing optimization methods: the multiplicative updates (MU) [61], PGM [20], and ADMM [29]. Each method is applied to the minimization problem (26) with the three types of loss function: the squared Euclidean distance (EUC), the KL divergence, and the Dual-IS divergence. Since PGM with Dual-IS has not been proposed in the literature, we omit a comparison with this specific combination. In the present simulations, the steady-state performance is independent of the initialization of $\boldsymbol{H}_k$, because the objective function has a unique minimizer due to its strict convexity and coercivity.

4.1. Performance in the coefficient errors for artificial data

We evaluate the performance of the proposed approach for artificial data. The objective function is different among the optimization methods, because MU and PGM can directly be applied to (26) only for $\lambda_1 = \lambda_2 = \lambda_3 = 0$. Therefore, the coefficient error (rather than the objective value) is adopted here as a performance measure for meaningful comparisons. The coefficient matrix $\boldsymbol{H}_k$ is initialized randomly with the i.i.d. uniform distribution $\mathcal{U}(0,1)$. We test two cases: test A: $(M, L, N) = (100, 20, 500)$, and test B: $(M, L, N) = (200, 40, 1000)$. Nonnegative matrices $\boldsymbol{W} \in \mathbb{R}_+^{M\times L}$ and $\boldsymbol{V} \in \mathbb{R}_+^{M\times N}$ are generated randomly from the i.i.d. uniform distributions: $w_{m,l} \sim \mathcal{U}(0,1)$ and $v_{m,n} \sim \mathcal{U}(0, 0.01)$. As the true solution $\boldsymbol{H}^* \in \mathbb{R}_+^{L\times N}$ is assumed componentwise-sparse and also row-sparse simultaneously, it is generated as follows: (i) the matlab function "SPRAND" is used to generate a matrix of density 0.5 (meaning that half of the components are zero) with its non-zero entries obeying the i.i.d. uniform distribution $\mathcal{U}(0, 10)$, and then (ii)

one out of every five rows of the generated matrix is forced to be a zero vector. A (noisy) data matrix to be factorized is generated by $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{H}^* + \boldsymbol{V}$. The number of total iterations is fixed to $10,000$ to evaluate the performance at the steady state as well as the transient behavior. The coefficient errors $\|\boldsymbol{H}_k - \boldsymbol{H}^*\|_{\mathrm{F}}$ are averaged over 100 independent trials.

In our preliminary experiments, it turned out that rough tuning was enough for the parameters λ_1 and γ, whereas fine tuning was necessary for λ_2 and λ_3 to obtain good performance. (The performance of the proposed method stayed nearly constant within a certain range around the optimal value for all the parameters.) For the proposed method, we first fix $\lambda_2 = 0.8$ and $\lambda_3 = 0.4$ tentatively, and find good parameters $\lambda_1 = 10^{-10}$ and $\gamma = 1$ by a coarse grid-search. The same parameter $\lambda_1 = 10^{-10}$ is used for ADMM.[3] We then fix those tuned parameters, and find the best parameters λ_2 and λ_3 by a fine grid-search for each of the proposed method and ADMM. The parameters are summarized in Table 2. We mention here that the parameter tuning is carried out for the first trial and the parameters are then fixed for the rest of the trials. As the proposed method converges to the global minimizer for any $\delta \in (0, 2)$, we let $\delta := 1.8$. For the PGM, the step size parameter is set to 1.0 for EUC, and to 0.1 for KL.

The results are illustrated in Figure 1, and are detailed in Tables 3 and 4, which show the average errors at the steady state with the 95% confidence intervals. We can see that the proposed method outperforms the existing methods when the KL or Dual-IS divergence is employed. We can also see that the proposed method and ADMM attain better performance than MU and PGM in all the cases. This implies that the regularizers work efficiently in estimation. The proposed method and ADMM attain comparable performance when the squared Euclidean distance is employed. This is because the proposed method uses no approximation in this case (the two methods therefore yield exactly the same solution (the global optimum) in theory).

To show the convergence behaviors, we plot in Figure 2 the learning curves in the coefficient errors for test A. We can see that the proposed method converges faster than ADMM and PGM on average. In the case of the squared Euclidean distance, in particular, although the steady-state performance is

[3]The parameter λ_1 controls the strength of the time-continuity regularizer, which is not important in the present experiment because $\boldsymbol{H}^*$ does not have such a continuous structure.

Table 2: Parameter settings for the artificial data.

	Divergence	λ_2	λ_3
Proposed	EUC	10^{-9}	10^{-3}
	KL	10^{-9}	4×10^{-3}
	Dual-IS	10^{-7}	4×10^{-3}
ADMM	EUC	10^{-9}	10^{-3}
	KL	10^{-9}	10^{-1}
	Dual-IS	10^{-7}	1

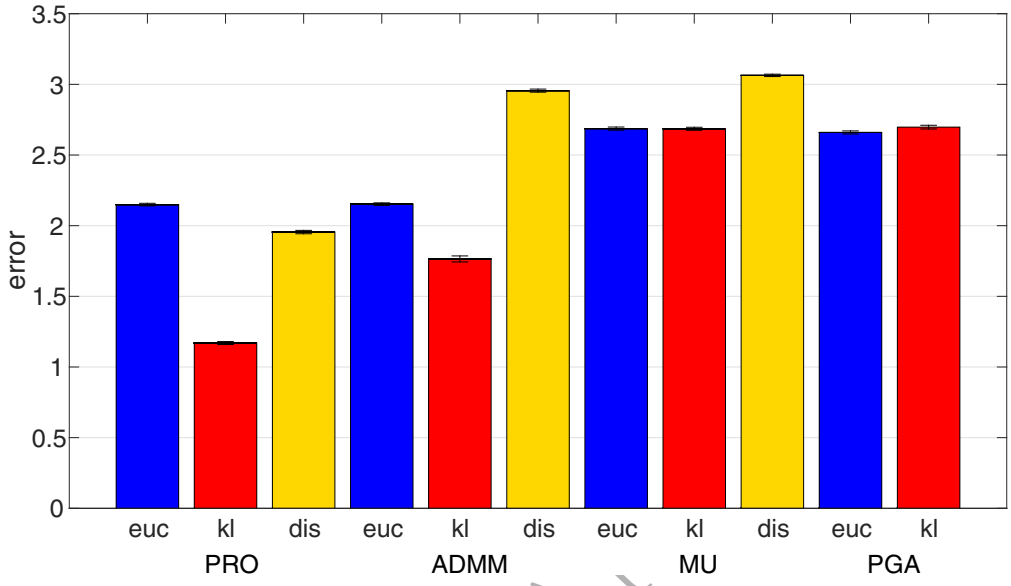
Table 3: Average steady-state errors for test A: $\boldsymbol{H} \in \mathbb{R}_+^{20\times 500}$

	Proposed	ADMM	MU	PGM
EUC	2.1494 ± 0.0090	2.1524 ± 0.0091	2.6869 ± 0.0111	2.6598 ± 0.0109
KL	**1.1695** ± 0.0105	1.7647 ± 0.0213	2.6857 ± 0.0107	2.6975 ± 0.0131
Dual-IS	1.9548 ± 0.0120	2.9556 ± 0.0109	3.0637 ± 0.0085	-

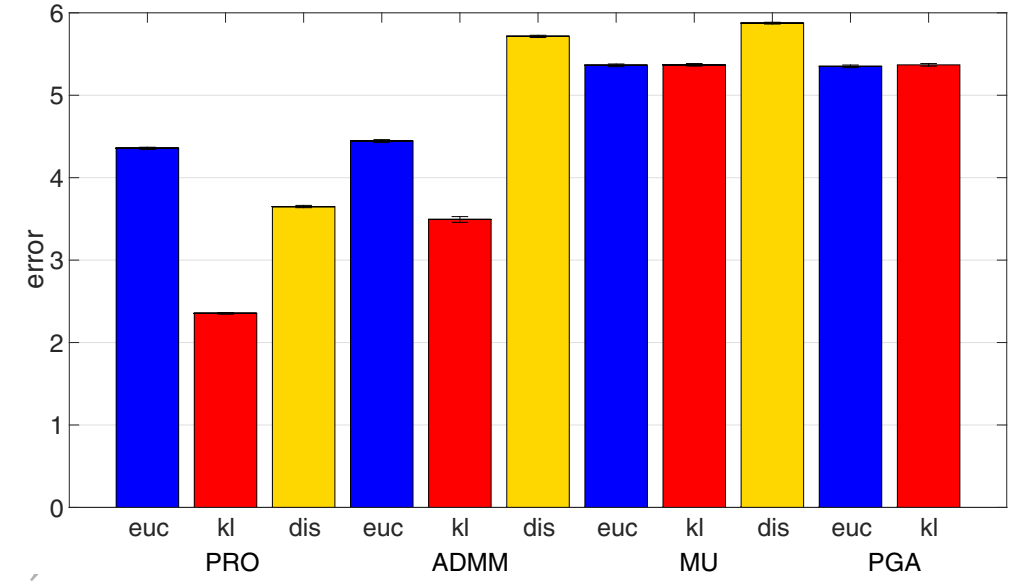
Table 4: Average steady-state errors for test B: $\boldsymbol{H} \in \mathbb{R}_+^{40\times 1000}$

	Proposed	ADMM	MU	PGM
EUC	4.3555 ± 0.0140	4.4439 ± 0.0178	5.3630 ± 0.0163	5.3510 ± 0.0157
KL	**2.3533** ± 0.0099	3.4931 ± 0.0349	5.3692 ± 0.0155	5.3688 ± 0.0155
Dual-IS	3.6499 ± 0.0139	5.7139 ± 0.0136	5.8747 ± 0.0109	-

the same between the proposed method and ADMM as explained above, the proposed method exhibits faster convergence than ADMM. MU exhibits fast initial convergence at the price of large errors at the steady state. We stress that the gain to ADMM comes from the algorithmic structure; i.e., the proposed method is free from auxiliary variables due to the use of Moreau envelope and is thus free from extra errors as well (see the first paragraph of Section 3).

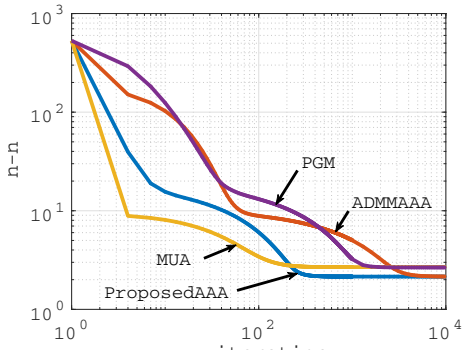


(a) Test A: $\boldsymbol{H} \in \mathbb{R}_{+}^{20\times 500}$

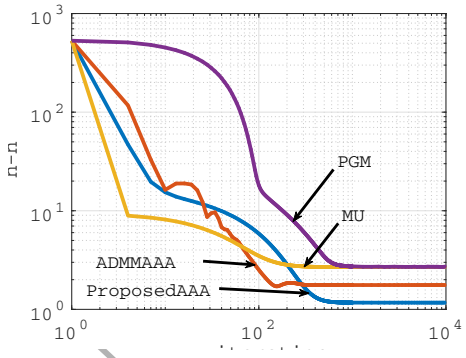


(b) Test B: $\boldsymbol{H} \in \mathbb{R}_{+}^{40\times 1000}$

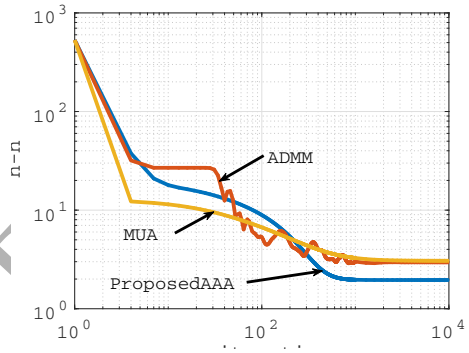
Figure 1: Average errors at the steady state.



(a) Euclidean distance



(b) KL divergence



(c) Dual-IS divergence

Figure 2: Convergence behaviors for test A: $\boldsymbol{H} \in \mathbb{R}_{+}^{20\times 500}$.

4.2. Application to Music Transcription

We apply our supervised NMF method to polyphonic music transcription. In this application, the matrix $\boldsymbol{Y}$ represents the spectrogram of the audio signal to be analyzed for each time bin, $\boldsymbol{W}$ contains the magnitude spectra of a wide range of pitches of musical instruments, and $\boldsymbol{H}$ is the activation matrix, each row of which represents the time variations of the sound pressure. We employ the standard (frame-based) performance-measures based on the MIREX [76]: the F-measure, the total errors, and the false alarms. We present a remark below on how to build the matrix $\boldsymbol{W}$ in this specific application.

Remark 4 (On training individual atoms in the dictionary). *Each column vector of $\boldsymbol{W}$ is generated simply by unsupervised NMF. A wide range of single isolated tones generated with different musical instruments are used as training audio signals. 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, the spectra corresponding to the isolated tones are acquired, and $\boldsymbol{W}$ is then obtained by normalizing those spectra. The single isolated tones for this generation process are available at, for instance, the RWC music database* [77]*, the MAPS database* [78]*, etc. See* [79, 80] *for other works that exploit prior information for NMF.*

Three test-data of polyphonic piano music are generated with several tones from the piano model "012PFNOF" contained in the RWC music database [77]; see Table 5 and Figure 3. The sampling frequency is 44.1 kHz. Each column of $\boldsymbol{Y}$ is computed with the short-time Fourier transform (STFT) using a Hamming window of length 23.22 ms with 50% overlap. The matrix $\boldsymbol{W}$ contains the spectra for 88 pitches of piano. Accordingly, the size of the matrices is given as follows: $M = 513$, $L = 88$, and $N = 1085$ (music data 1), $N = 1290$ (music data 2), $N = 1951$ (music data 3). The test data contain the tones that are also used to generate the dictionary $\boldsymbol{W}$. The output matrix $\boldsymbol{H}$ is binarized for the music transcription task with the threshold 5% of the maximal value of the matrix $\boldsymbol{H}$ itself.[4]

For the proposed method, the parameters λ_1 and γ are set, for all the music data, to $\lambda_1 = 10^{-5}$ and $\gamma = 1.2$, which gave reasonable performance

[4]Choosing the threshold value in this way worked slightly better in our experiments than using a fixed threshold value for each method.

Table 5: Descriptions of the music data.

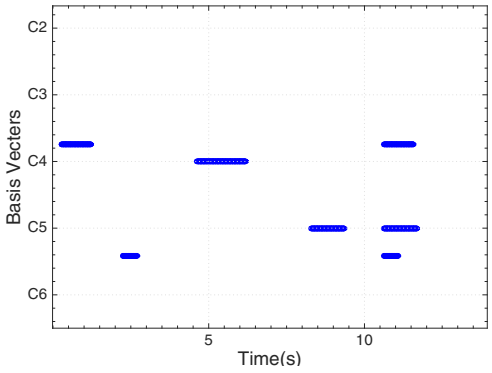
	# sources	components	duration
music data 1	4	A3, C4, C5, F5	$\approx$ 13 sec.
music data 2	4	G#3, F#4, F#5, D6	$\approx$ 15 sec.
music data 3	3	C4, E4, G4	$\approx$ 23 sec.

for music data 1. One may regard that those parameters are tuned with the training data (music data 1) and then applied to the test data (music data 2 and 3). The parameter λ_1 controls the degree of time continuity of the solution; the larger it is, the smaller the time-variations of each row vector of $\boldsymbol{H}_k$ become. The parameters λ_2 and λ_3 are tuned with a grid-search for each music data so that the best performance in total errors is obtained. We let $\delta := 1.8$ as in Section 4.1.

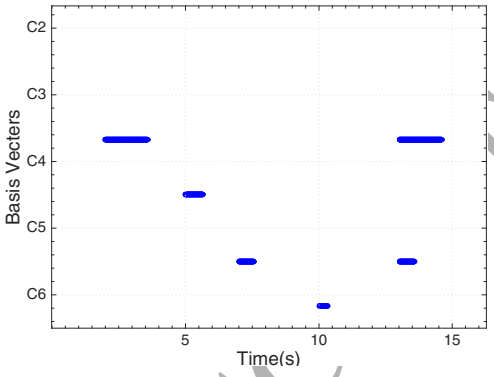
Tables 6–8 summarize the parameter settings and the results for each music data. Our observations are summarized below.

1. The proposed method significantly outperforms the existing methods. As in Section 4.1, the gain compared to MU and PGM comes from the use of multiple regularizers, while the gain compared to ADMM comes from no use of auxiliary variables (see the first paragraph of Section 3).
2. The KL and Dual-IS divergences offer better performance than the squared Euclidean distance.
3. The use of the regularizers ameliorates the performance.
4. For the proposed method and ADMM, Dual-IS gives slightly better performance than KL, while KL gives much better performance than Dual-IS for MU. This suggests that the Dual-IS divergence works well when it is employed with the regularizers studied in the present work.

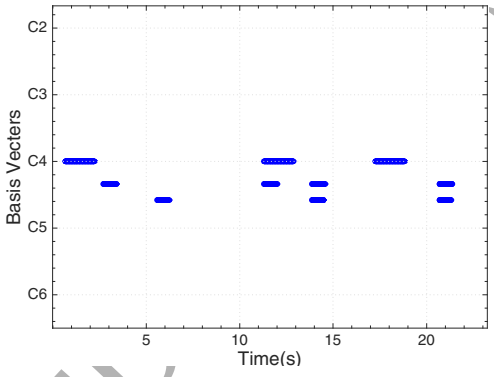
The music transcription results for music data 1 are illustrated in Figures 5 and 6. (We had similar results for the other music data as well.) The blue lines indicate correct transcription, the green lines indicate missed errors, and the red crosses indicate false alarms. The results are consistent with those in Tables 6–8.



(a) Music data 1



(b) Music data 2



(c) Music data 3

Figure 3: Ground truth for the music data.

Table 6: Parameter settings and the results for music data 1.

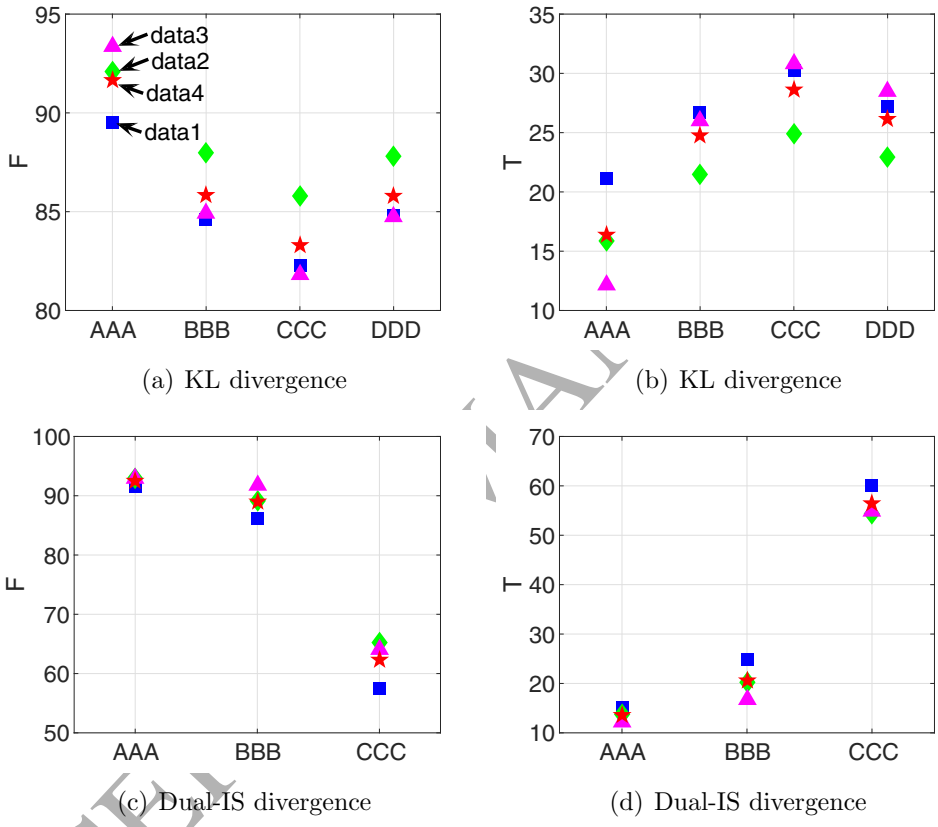
Method	Loss Function	Parameter		Results		
		λ_2	λ_3	$\mathcal{F}$	$\mathcal{E}_{\text{tot}}$	$\mathcal{E}_{\text{fals}}$
Proposed	EUC	0.8	0.4	83.64	29.14	4.19
	KL	0.8	0.4	89.50	21.14	13.50
	DIS	0.02	0.04	**91.62**	**15.10**	8.35
ADMM	EUC	0.8	0.4	83.86	28.92	4.64
	KL	0.8	0.4	84.63	26.71	0.22
	DIS	0.2	0.04	**86.27**	**24.72**	3.09
MU	EUC	–	–	**83.50**	**29.36**	4.19
	KL	–	–	82.31	30.24	1.10
	DIS	–	–	57.58	60.04	1.55
PGA	EUC	–	–	83.72	29.14	4.64
	KL	–	–	**84.80**	**27.15**	3.53

Table 7: Parameter settings and the results for music data 2.

Method	Loss Function	Parameter		Results		
		λ_2	λ_3	$\mathcal{F}$	$\mathcal{E}_{\mathrm{tot}}$	$\mathcal{E}_{\mathrm{fals}}$
Proposed	EUC	0.8	0.2	82.87	30.02	6.04
	KL	0.8	0.2	92.07	15.89	8.17
	DIS	0.2	0.08	**93.00**	**13.69**	4.64
ADMM	EUC	0.8	0.2	83.43	29.13	6.04
	KL	0.8	0.4	87.99	21.49	0.89
	DIS	0.08	0.04	**89.09**	**20.25**	9.41
MU	EUC	–	–	82.49	30.55	7.10
	KL	–	–	**85.80**	**24.87**	1.07
	DIS	–	–	65.20	54.35	6.93
PGA	EUC	–	–	81.96	32.68	11.37
	KL	–	–	**87.80**	**22.91**	9.24

Table 8: Parameter settings and the results for music data 3.

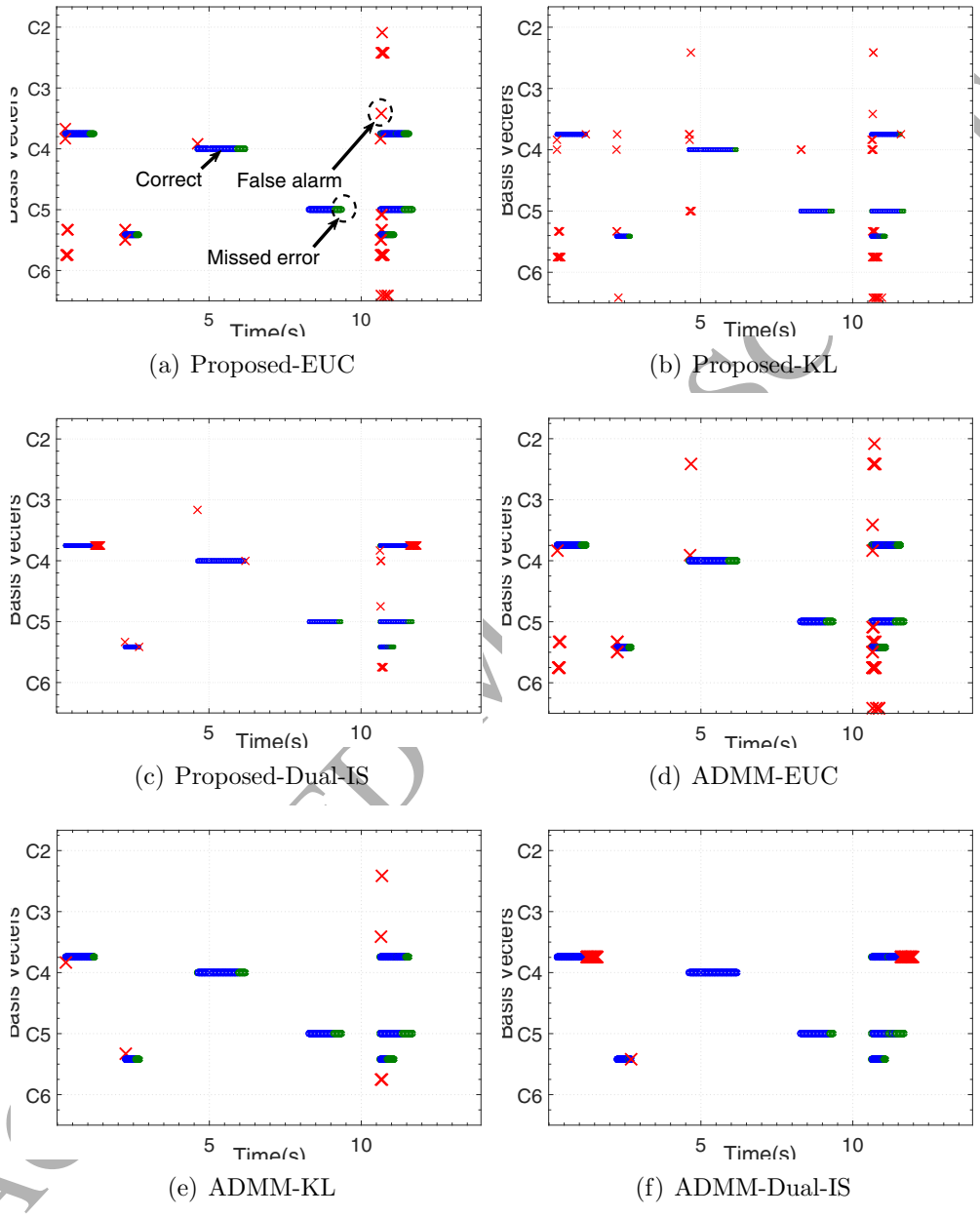
Method	Loss Function	Parameter		Results		
		λ_2	λ_3	$\mathcal{F}$	$\mathcal{E}_{\mathrm{tot}}$	$\mathcal{E}_{\mathrm{fals}}$
Proposed	EUC	0.8	0.04	77.15	37.64	4.94
	KL	1.2	0.05	**93.37**	**12.14**	6.32
	DIS	0.2	0.04	92.91	**12.14**	2.23
ADMM	EUC	0.8	0.04	77.38	37.54	5.55
	KL	0.8	0.04	84.94	26.03	0.30
	DIS	0.1	0.4	**91.80**	**16.65**	14.43
MU	EUC	–	–	76.75	38.45	5.75
	KL	–	–	**81.81**	**30.78**	0.91
	DIS	–	–	64.12	54.89	4.84
PGA	EUC	–	–	74.82	42.38	11.50
	KL	–	–	**84.76**	**28.46**	8.98



(a) KL divergence (b) KL divergence

(c) Dual-IS divergence (d) Dual-IS divergence

Figure 4: F-measure and total errors.

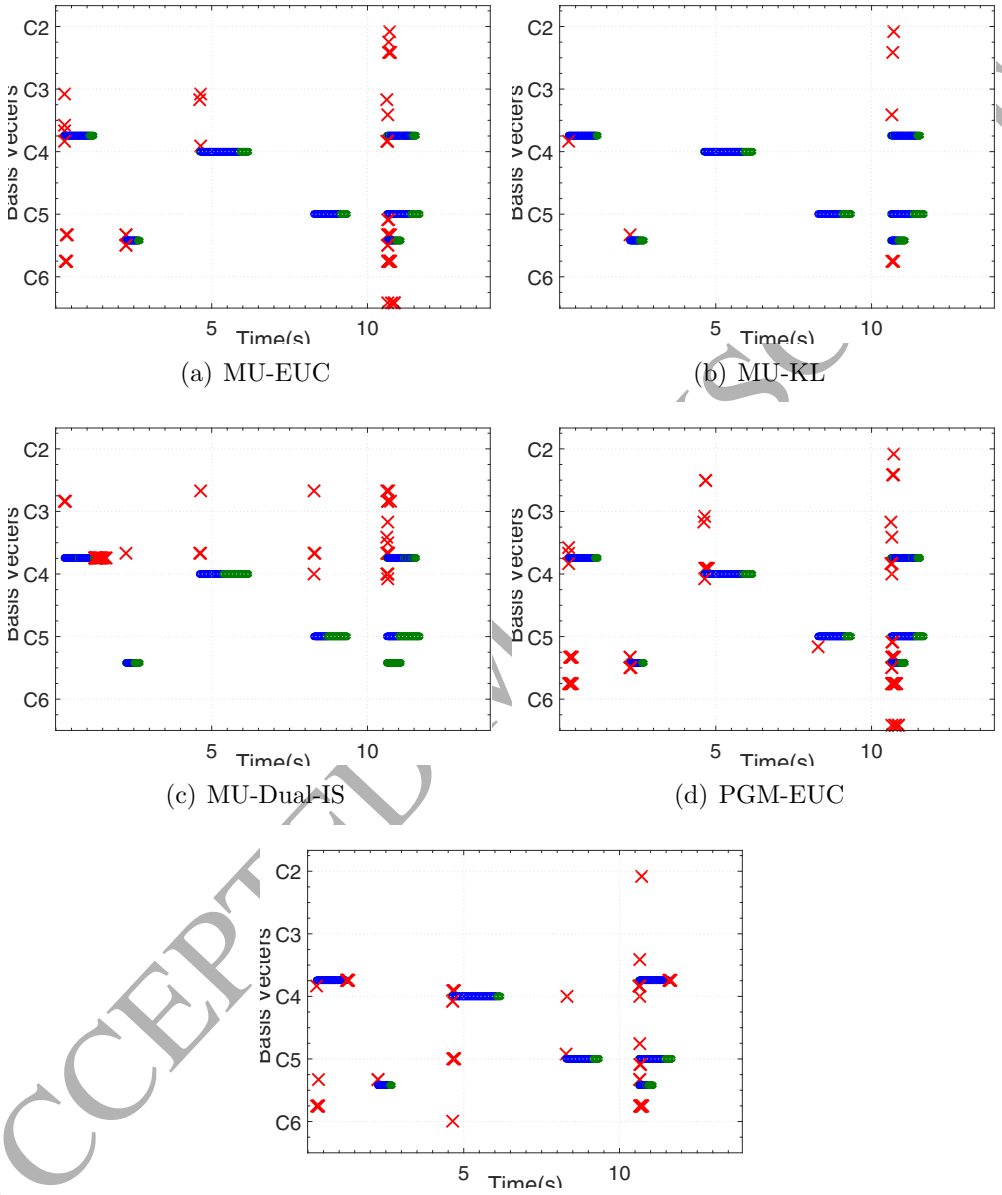


(a) Proposed-EUC (b) Proposed-KL

(c) Proposed-Dual-IS (d) ADMM-EUC

(e) ADMM-KL (f) ADMM-Dual-IS

Figure 5: Music transcription results of the proposed method and ADMM for music data 1.



(a) MU-EUC (b) MU-KL

(c) MU-Dual-IS (d) PGM-EUC

(e) PGM-KL

Figure 6: Music transcription results of MU and PGM for music data 1.

5. Conclusion

We proposed an efficient supervised NMF approach using the Moreau envelope, a smooth approximation, of the KL/Dual-IS divergence. The useful properties of the Moreau envelope and the proximity operator (both of which are closely related to each other) were presented. The supervised NMF problem was cast as minimization of the loss function penalized by four terms: (i) the time continuity, (ii) the indicator function ensuring the nonnegativity, (iii) the basis-vector selector, and (iv) the sparsity-promoting regularizer. A closed-form expression of the proximity operator of the sum of the three non-differentiable terms (ii)–(iv) was derived. The minimization problem can therefore be solved numerically by the proximal forward-backward splitting method. The proposed approach requires no auxiliary variable, and is therefore free from the extra errors in addition that it is memory efficient. The simulation results showed the efficacy of the proposed method in an application to polyphonic music transcription. It should be remarked again that the challenging task of determining the exact number of sources prior to decomposition is unnecessary in the proposed supervised approach. As demonstrated already in [25], the proposed method is robust against possible imperfection in $\boldsymbol{W}$. Finally, although our focus was solely on the supervised approach in the present study to build a theoretically sound method, the developed techniques will also be useful in the unsupervised NMF framework.

Acknowledgment: This work was supported by the Support Center for Advanced Telecommunications Technology Research (SCAT) and JSPS Grants-in-Aid (15K06081, 15K13986, 15H02757).

[1] D. D. Lee, H. S. Seung, Learning the parts of objects by nonnegative matrix factorization, Nature 401 (1999) 788–791.

[2] D. R. Hunter, K. Lange, A tutorial on MM algorithms, The American Statistician 58 (1) (2004) 30–37.

[3] L. B. Thomas, Rank factorization of nonnegative matrices (A. Berman), SIAM Rev. 16 (1974) 393–394.

[4] S. L. Campbell, G. D. Poole, Computing nonnegative rank factorizations, Linear Algebra Appl. 35 (1981) 175–182.

[5] J.-C. Chen, The nonnegative rank factorizations of nonnegative matrices, Linear Algebra Appl. 62 (1984) 207–217.

[6] J. Cohen, U. Rothblum, Nonnegative ranks, decompositions and factorizations of nonnegative matrices, Linear Algebra Appl. 190 (1993) 149–168.

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

[8] T. Hofmann, Probabilistic latent semantic analysis, in: Proc. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, Berkeley, CA, 1999, pp. 50–57.

[9] I. S. Dhillon, S. Sra, Generalized nonnegative matrix approximations with bregman divergences, in: Neural Information Processing Systems (NIPS), 2005, pp. 283–290.

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

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

[12] 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.

[13] T. Virtanen, Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria, IEEE Trans. Audio, Speech, and Language Processing 15 (3) (2007) 1066–1074.

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

[15] D. L. Donoho, V. C. Stodden, When does non-negative matrix factorization give a correct decomposition into parts?, in: Advances in Neural Information Processing Systems (NIPS), Vol. 16, MIT Press, Cambridge, MA, USA, 2003, pp. 1141–1148.

[16] S. Moussaoui, D. Brie, J. Idier, Non-negative source separation: Range of admissible solutions and conditions for the uniqueness of the solution, in: Proc. ICASSP, Vol. 5, 2005, pp. 289–292.

[17] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, S. H. Jensen, Theorems on positive data: On the uniqueness of NMF, Comput. Intell. Neurosci. 2008.

[18] K. Huang, N. D. Sidiropoulos, A. Swami, Non-negative matrix factorization revisited: uniqueness and algorithm for symmetric decomposition, IEEE Trans. Signal Processing 62 (1) (2014) 211–224.

[19] S. A. Vavasis, On the complexity of nonnegative matrix factorization, Siam J. Optim. 20 (3) (2009) 1364–1377.

[20] C.-J. Lin, Projected gradient methods for non-negative matrix factorization, Neural Computation 19 (2007) 2756–2779.

[21] D. Kim, S. Sra, I. S. Dhillon, Fast projection-based methods for the least squares nonnegative matrix approximation problem, Statistical Analysis and Data Mining 1 (2008) 38–51.

[22] N. Gillis, Nonnegative matrix factorization: Complexity, algorithms and applications", Ph.D. thesis, Université catholique de Louvain (2011).

[23] R. Tandon, S. Sra, Sparse nonnegative matrix approximation: new formulations and algorithms, Tech. Rep. 193, Max Planck Institute for Biological Cybernetics (2010).

[24] E. Esser, M. Möller, S. Osher, G. Sapiro, J. Xin, A convex model for nonnegative matrix factorization and dimensionality reduction on physical space, IEEE Trans. Image Processing 21 (7) (2012) 3239–3252.

[25] Y. Morikawa, M. Yukawa, A sparse optimization approach to supervised NMF based on convex analytic method, in: Proc. IEEE ICASSP, 2013, pp. 6078–6082.

[26] D. L. Sun, C. Févotte, Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence, in: Proc. ICASSP, 2014, pp. 6201–6205.

[27] F. Yanez, F. Bach, Primal-dual algorithms for non-negative matrix factorization with the kullback-leibler divergence, arXiv:1412.1788 [cs.LG], <hal-01079229> (2014).

[28] R. Zdunek, Alternating direction method for approximating smooth feature vectors in nonnegative matrix factorization, in: Proc. IEEE MLSP, 2014.

[29] H. Kagami, M. Yukawa, Supervised nonnegative matrix factorization with dual-Itakura-Saito and Kullback-Leibler divergences for music transcription, in: Proc. European Signal Processing Conference (EUSIPCO), 2016, pp. 1138–1142.

[30] A. Ben-Tal, M. Teboulle, A smoothing technique for nondifferentiable optimization problems, in: Optimization, Lecture Notes in Math., Vol. 1405, Springer-Verlag, Berlin, 1989, pp. 1–11.

[31] C. Lemaréchal, C. Sagastizábal, Practical aspects of the MoreauYosida regularization: Theoretical preliminaries, SIAM J. Optim. 7 (1997) 367–385.

[32] Y. Nesterov, Smooth minimization of nonsmooth functions, Math. Program. 103 (2005) 127–152.

[33] J. C. Duchi, P. L. Bartlett, M. J. Wainwright, Randomized smoothing for stochastic optimization, SIAM J. Optim. 22 (2) (2012) 674–701.

[34] P. L. Lions, B. Mercier, Splitting algorithms for the sum of two nonlinear operators, SIAM J. Numer. Anal. 16 (1979) 964–979.

[35] P. L. Combettes, V. R. Wajs, Signal recovery by proximal forward-backward splitting, SIAM Multiscale Model. Simul. 4 (2005) 1168–1200.

[36] H. Raguet, J. Fadili, G. Peyré, Generalized forward-backward splitting, Arxiv preprint arXiv:1108.4404v3 [math.OC].

[37] R. Glowinski, A. Marrocco, Sur l'approximation, par elements finis d'ordre un, et la resolution, par penalisation-dualité, d'une classe de problems de dirichlet non lineares, Revue Francaise d'Automatique, Informatique, et Recherche Opérationelle 9 (1975) 41–76.

[38] D. Gabay, B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximations, Computers and Mathematics with Applications 2 (1976) 17–40.

[39] A. A. Goldstein, Convex programming in Hilbert space, Bull. Amer. Math. Soc. 70 (1964) 709–710.

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

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

[42] Y.-X. Wang, Y.-J. Zhang, Nonnegative matrix factorization: A comprehensive review, IEEE Trans. Knowledge and Data Engineering 25 (6) (2013) 1336–1353.

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

[44] I. Yamada, M. Yukawa, 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, Springer, 2011, pp. 345–390.

[45] H. H. Bauschke, J. M. Borwein, Legendre functions and the method of random Bregman projections, J. Convex Anal. 4 (1997) 27–67.

[46] H. H. Bauschke, J. M. Borwein, P. L. Combettes, Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces, Commun. Contemp. Math. 3 (2001) 615–647.

[47] E. Resmerita, On total convexity, Bregman projections and stability in Banach spaces, J. Convex Anal. 11 (1) (2004) 1–16.

[48] H. H. Bauschke, P. L. Combettes, Construction of best Bregman approximations in reflexive Banach spaces, in: Proc. Amer. Math. Soc., Vol. 131, 2003, pp. 3757–3766.

[49] S. S. Reich, S. Sabach, Existence and approximation of fixed points of Bregman firmly nonexpansive mappings in reflexive Banach spaces, in: Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer, 2011, pp. 301–316.

[50] L. M. Bregman, The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming, USSR Comput. Math. Math. Phys. 7 (3) (1967) 210–217.

[51] S. Amari, H. Nagaoka, Methods of Information Geometry, Oxford University Press, New York, NY, USA, 2000.

[52] S. Amari, α-divergence is unique, belonging to both f-divergence and bregman divergence classes, IEEE Trans. Inform. Theor. 55 (2009) 4925–4931.

[53] A. Basu, I. R. Harris, N. L. Hjort, M. C. Jones, Robust and efficient estimation by minimising a density power divergence, Biometrika 85 (3) (1998) 549–559.

[54] S. Eguchi, Y. Kano, Robustifying maximum likelihood estimation, Tech. rep., Institute of Statistical Mathematics (Jun. 2001).

[55] M. Minami, S. Eguchi, Robust blind source separation by beta-divergence, Neural Comput. 14 (2002) 1859–1886.

[56] F. Liese, I. Vajda, Convex Statistical Distances, Vol. 95 of Teubner-Texte zur Mathematik, Leipzig: Teubner, 1987.

[57] J. Lafferty, Additive models, boosting, and inference for generalized divergences, in: Proc. Annual Conference on Computational Learning Theory (COLT), 1999.

[58] A. Cichocki, S. Amari, Families of Alpha-Beta- and Gamma- divergences: Flexible and robust measures of similarities, Entropy 12 (2010) 1532–1568.

[59] R. Hennequin, B. David, R. Badeau, Beta-divergence as a subclass of Bregman divergence, IEEE Signal Processing Letters 18 (2) (2011) 83–86.

[60] C. Févotte, A. T. Cemgil, Nonnegative matrix factorisations as probabilistic inference in composite models, in: Proc. European Signal Processing Conference (EUSIPCO), 2009, pp. 1913–1917.

[61] A. Cichocki, S. Cruces, S. i. Amari, Generalized Alpha-Beta divergences and their application to robust nonnegative matrix factorization, Entropy 13 (2011) 134–170.

[62] J. J. Moreau, Fonctions convexes duales et points proximaux dans un espace hilbertien, C. R. Acad. Sci. Paris Ser. A Math. 255 (1962) 2897–2899.

[63] J. J. Moreau, Proximité et dualité dans un espace hilbertien, Bull. Soc. Math. France 93 (1965) 273–299.

[64] P. L. Combettes, J.-C. Pesquet, Proximal splitting methods in signal processing, in: Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer, 2011, pp. 185–212.

[65] G. Mysore, A non-negative framework for joint modeling of spectral structure and temporal dynamics in sound mixtures, Ph.D. thesis, Stanford University (2010).

[66] P. L. Combettes, C. L. Müller, Perspective functions: Proximal calculus and applications in high-dimensional statistics, J. Math. Anal. Appl.

[67] W. R. Mann, Mean value methods in iteration, Proc. Amer. Math. Soc. 4 (1953) 506–510.

[68] M. A. Krasnosel'skiĭ, Two remarks on the method of successive approximations, Uspekhi Mat. Nauk 10 (1(63)) (1955) 123–127, (in Russian).

[69] G. Grindlay, D. P. W. Ellis, Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments, IEEE Journal of Selected Topics in Signal Processing 5 (6) (2011) 1159–1169.

[70] A. Dessein, A. Cont, G. Lemaitre, Real-time detection of overlapping sound events with non-negative matrix factorization, in: Matrix Information Geometry, Springer, 2012, pp. 341–371.

[71] D. Guillamet, J. Vitriá, B. Schiele, Introducing a weighted nonnegative matrix factorization for image classification, Pattern Recognit. Lett. 24 (14) (2003) 2447–2454.

[72] E. Vincent, Musical source separation using time-frequency source priors, IEEE Trans. Audio, Speech, Lang. Process. 14 (1) (2006) 91–98.

[73] A. Cont, Realtime multiple pitch observation using sparse non-negative constraints, in: Proc. Int. Symp. Music Inform. Retrieval, 2006, pp. 206–212.

[74] G. Peyré, Non-negative sparse modeling of textures, in: Proc. Scale Space Variational Methods Comput. Vis., 2007, pp. 628–639.

[75] P. Smaragdis, B. Raj, M. Shashanka, Supervised and semisupervised separation of sounds from single-channel mixtures, in: Proc. LVA/ICA, 2010, pp. 140–148.

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

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

[78] V. Emiya, R. Badeau, B. David, Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle, IEEE Trans. Audio, Speech and Language Processing 18 (6) (2010) 1643–1654.

[79] S. Ewert, M. Müller, Using score-informed constraints for NMF-based source separation, in: Proc. IEEE ICASSP, 2012, pp. 129–132.

[80] F. Weninger, C. Kirst, B. Schuller, H. J. Bungartz, A discriminative approach to polyphonic piano note transcription using supervised non-negative matrix factorization, in: Proc. IEEE ICASSP, 2013, pp. 6–10.