Nonnegative matrix factorization with α-divergence

Andrzej Cichocki [a], Hyekyoung Lee [b], Yong-Deok Kim [b], Seungjin Choi [b,*]

[a] *Laboratory for Advanced Brain Signal Processing*
Brain Science Institute, RIKEN
2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan

[b] *Department of Computer Science*
Pohang University of Science and Technology
San 31 Hyoja-dong, Nam-gu, Pohang 790-784, Korea

Abstract

Nonnegative matrix factorization (NMF) is a popular technique for pattern recognition, data analysis, and dimensionality reduction, the goal of which is to decompose nonnegative data matrix $\boldsymbol{X}$ into a product of basis matrix $\boldsymbol{A}$ and encoding variable matrix $\boldsymbol{S}$ with both $\boldsymbol{A}$ and $\boldsymbol{S}$ allowed to have only nonnegative elements. In this paper we consider Amari's α-divergence as a discrepancy measure and rigorously derive a multiplicative updating algorithm (proposed in our recent work) which iteratively minimizes the α-divergence between $\boldsymbol{X}$ and $\boldsymbol{AS}$. We analyze and prove the monotonic convergence of the algorithm using auxiliary functions. In addition, we show that the same algorithm can be also derived using Karush-Kuhn-Tucker (KKT) conditions as well as the projected gradient. We provide two empirical study for image denoising and EEG classification, showing the interesting and useful behavior of the algorithm in cases where different values of α ($\alpha = 0.5, 1, 2$) are used.

Key words: α-divergence, Multiplicative updates, Nonnegative matrix factorization, Projected gradient

* Corresponding author. Tel.: +82-54-279-2259; Fax: +82-54-279-2299
Email: cia@brain.riken.jp (A. Cichocki),
leehk@postech.ac.kr (H. Lee)
karma13@postech.ac.kr (Y. -D. Kim)
seungjin@postech.ac.kr (S. Choi)
URL: http://www.postech.ac.kr/∼seungjin (S. Choi)

1 Introduction

Nonnegative matrix factorization (NMF) is one of widely-used multivariate data analysis methods (Paatero and Tapper, 1997; Lee and Seung, 1999, 2001), which has many potential applications in pattern recognition and machine learning. These applications include face recognition (Li et al., 2001), document clustering (Xu et al., 2003; Shahnaz et al., 2006), sound classification (Cho and Choi, 2005), medical imaging (Lee et al., 2001; Ahn et al., 2004), audio processing (Smaragdis and Brown, 2003; Kim and Choi, 2006), bioinformatics (Brunet et al., 2004), and so on.

Suppose that N observed data points, $\{\boldsymbol{x}_t\}$, $t = 1, \ldots, N$ are available. Denote the data matrix by $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{m \times N}$. NMF seeks a decomposition of the nonnegative data matrix $\boldsymbol{X}$ that is of the form:

$$\boldsymbol{X} \approx \boldsymbol{A}\boldsymbol{S}, \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ contains basis vectors in its columns and $\boldsymbol{S} \in \mathbb{R}^{n \times N}$ is the associated encoding variable matrix. Both matrices $\boldsymbol{A}$ and $\boldsymbol{S}$ are restricted to have only nonnegative elements in the decomposition.

Various error measures for the factorization (1) with nonnegativity constraints, were considered, including sparseness constraints (Hoyer, 2004), Csiszár's divergence (Cichocki et al., 2006b,c), Bregman divergence (Dhillon and Sra, 2006), and a generalized divergence measure Kompass (2007). and the convergence issue was recently studied (Lin, 2007). Two widely-used error measures that were considered in (Lee and Seung, 2001) are summarized:

(1) (**LS**) Least squares (LS) criterion which employs the Euclidean distance between the data matrix $\boldsymbol{X}$ and the model $\boldsymbol{A}\boldsymbol{S}$ is given by

$$\mathcal{E}_1 = \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{S}\|^2 = \sum_{i,j} \left[X_{ij} - [\boldsymbol{A}\boldsymbol{S}]_{ij}\right]^2. \tag{2}$$

(2) (**I-divergence, KL-divergence**) I-divergence between $\boldsymbol{X}$ and $\boldsymbol{A}\boldsymbol{S}$ is given by

$$\mathcal{E}_2 = \sum_{i,j} \left[X_{ij} \log \frac{X_{ij}}{[\boldsymbol{A}\boldsymbol{S}]_{ij}} - X_{ij} + [\boldsymbol{A}\boldsymbol{S}]_{ij}\right]. \tag{3}$$

The minimization of the objective functions described above, should be done with nonnegativity constraints for both $\boldsymbol{A}$ and $\boldsymbol{S}$. Multiplicative updating is an efficient way in such a case, since it can easily preserve nonnegativity constraints at each iteration. Multiplicative updating algorithms for NMF associated with these two objective functions are given as follows:

(1) (**LS**) A local minimum of the objective function (2) is computed by the LS multiplicative algorithm that has the form

$$S_{ij} \leftarrow S_{ij} \frac{\left[\boldsymbol{A}^T \boldsymbol{X}\right]_{ij}}{\left[\boldsymbol{A}^T \boldsymbol{A} \boldsymbol{S}\right]_{ij}}, \quad A_{ij} \leftarrow A_{ij} \frac{\left[\boldsymbol{X} \boldsymbol{S}^T\right]}{\left[\boldsymbol{A} \boldsymbol{S} \boldsymbol{S}^T\right]_{ij}}. \tag{4}$$

(2) (**I-divergence**) In the case of I-divergence-based objective function (3), its minimum is found by the multiplicative updating algorithm that is of the form

$$S_{ij} \leftarrow S_{ij} \frac{\sum_k \left[A_{ki} X_{kj} / \left[\boldsymbol{A} \boldsymbol{S}\right]_{kl}\right]}{\sum_l A_{li}}, \quad A_{ij} \leftarrow A_{ij} \frac{\sum_k \left[S_{jk} X_{ik} / \left[\boldsymbol{A} \boldsymbol{S}\right]_{ik}\right]}{\sum_l S_{jl}}. \tag{5}$$

Recently different error measures such as Csiszár's f-divergences, Amari's α-divergences, and Bregman divergences, were considered in the context of NMF (Cichocki et al., 2006b,c; Dhillon and Sra, 2006). Multiplicative NMF algorithms were proposed in (Cichocki et al., 2006b,a,c; Dhillon and Sra, 2006), considering Amari's α-divergence (Amari, 1985; Zhu and Rohwer, 1995) which is a special instance of Csiszár's f-divergence (Ali and Silvey, 1966; Csiszár, 1974). In this paper we derive an α-divergence-based NMF multiplicative algorithm in a different way as well as in a rigorous manner, with proving the monotonic local convergence of the algorithm using auxiliary functions. We also show that the same algorithm can be derived using Karush-Kuhn-Tucker (KKT) conditions as well as the projected gradient. Our contribution is primarily in the derivation of a generic multiplicative algorithm, its monotonic convergence, and alternative views when α-divergence is used as a discrepancy measure in the context of NMF.

We provide two numerical experiments. In the first experiment we apply our algorithm to the task of image denoising when three different noises (pepper, salt, and pepper-salt) were involved. We show how our algorithm with different values of α ($\alpha = 0.5, 1, 2$) behaves for different noise types. In the second experiment, we apply our algorithm to EEG data, demonstrating the useful behavior as a feature extractor for EEG classification and also investigating the classification performance for several different values of α.

2 Amari's α-divergence

Let us consider two unnormalized distributions $p(x)$ and $q(x)$ associated with a random variable x. Kullback-Leibler (KL) divergence is defined by

$$KL[p||q] = \int p \log \frac{p}{q} \, d\mu - \int (p - q) \, d\mu, \tag{6}$$

where μ is the Lebesque measure which is a shorthand notation for $\mu(dx)$. Note that the second term $\int(p-q)d\mu$ disappears when p and q are normalized distributions, i.e., $\int p\,d\mu = \int q\,d\mu = 1$. KL-divergence in (6) is often referred to as *I-divergence*.

The α-divergence (Amari, 1985) is a parametric family of divergence functional, including several well-known divergence measures as its special cases.

Definition 1 (α-divergence) *The α-divergence is defined by*

$$D_\alpha[p||q] = \frac{1}{\alpha(1-\alpha)} \int \alpha p + (1-\alpha)q - p^\alpha q^{1-\alpha}\, d\mu, \tag{7}$$

where $\alpha \in (-\infty, \infty)$.

As in KL divergence, α-divergence is zero if $p = q$ and positive otherwise. This property follows from the fact that α divergence (7) is convex with respect to p and q. The α-divergence includes KL-divergence, Hellinger divergence, and χ^2-divergence (Pearson's distance), as its special cases, which is summarized below.

- The α-divergence in (7) is often represented by

$$D_\beta[p||q] = \frac{4}{1-\beta^2} \int \frac{1-\beta}{2}p + \frac{1+\beta}{2}q - p^{\frac{1-\beta}{2}} q^{\frac{1+\beta}{2}}\, d\mu, \tag{8}$$

  which is obtained by setting $\alpha = \frac{1-\beta}{2}$ and $1-\alpha = \frac{1+\beta}{2}$ in (7).
- As α approaches 0, α-divergence specializes to KL-divergence from q to p:

$$\lim_{\alpha\to 0} D_\alpha[p||q] = KL[q||p]. \tag{9}$$

- For $\alpha = \frac{1}{2}$, α-divergence specializes to Hellinger divergence:

$$D_{\alpha=\frac{1}{2}}[p||q] = 2\int \left(\sqrt{p} - \sqrt{q}\right)^2\, d\mu. \tag{10}$$

- As α approaches 1, α-divergence specializes to KL divergence from p to q:

$$\lim_{\alpha\to 1} D_\alpha[p||q] = KL[p||q]. \tag{11}$$

- For $\alpha = 2$, α-divergence is identical to χ^2-divergence:

$$D_{\alpha=2}[p||q] = \frac{1}{2}\int \frac{(p-q)^2}{q}\, d\mu. \tag{12}$$

The α-divergence belongs to a family of convex divergence measures which is known as *Csiszár's f-divergence*, called sometimes also *Ali-Silvey divergence* (Ali and Silvey, 1966; Csiszár, 1974).

Definition 2 (f-divergence) *The Csiszár's f-divergence is defined by*

$$I_f[p||q] = \int p\, f\left(\frac{q}{p}\right) d\mu, \tag{13}$$

where $f(z)$ is a convex function, $f : [0, \infty) \mapsto (-\infty, \infty]$, which is continuous at 0, satisfying $f(1) = 0$ and $f'(1) = 0$.

Examples of Ciszár's f-divergences are (Cichocki et al., 2006b):

- When $f(z) = z - \log z - 1$, f-divergence specializes to KL divergence: $I_f[p||q] = KL[p||q]$.
- When $f(z) = \frac{1}{\alpha(1-\alpha)} \{\alpha + (1-\alpha)z - z^{1-\alpha}\}$, f-divergence specializes to α-divergence: $I_f[p||q] = D_\alpha[p||q]$.

Note that the Csiszár's f-divergences are a large class, but they do not include L_2 distance, $\int (p-q)^2\, d\mu$ (Cichocki et al., 2006b). Note also that f-divergences are always nonnegative and zero if $p = q$.

3 Algorithm derivation

We consider the objective function for α-NMF, that is based on the α-divergence between $\boldsymbol{X}$ and $\boldsymbol{AS}$, which is given by

$$D_\alpha[\boldsymbol{X}||\boldsymbol{AS}] = \frac{1}{\alpha(1-\alpha)} \sum_{i=1}^{m} \sum_{j=1}^{N} \alpha X_{ij} + (1-\alpha)[\boldsymbol{AS}]_{ij} - X_{ij}^{\alpha} [\boldsymbol{AS}]_{ij}^{1-\alpha}. \tag{14}$$

As in (Lee and Seung, 2001), we introduce an auxiliary function that is used in convergence analysis as well as in algorithm derivation.

Definition 3 (Auxiliary function) *A function $G(\boldsymbol{S}, \widetilde{\boldsymbol{S}})$ is said to be an auxiliary function for $F(\boldsymbol{S})$ if the following two conditions are satisfied:*

$$G(\boldsymbol{S}, \boldsymbol{S}) = F(\boldsymbol{S}), \text{ and } G(\boldsymbol{S}, \widetilde{\boldsymbol{S}}) \geq F(\boldsymbol{S}), \text{ for all } \widetilde{\boldsymbol{S}}.$$

Lemma 1 *The function*

$$G(\boldsymbol{S}, \widetilde{\boldsymbol{S}}) = \frac{1}{\alpha(1-\alpha)} \sum_{i,j,k} X_{ij} \zeta_{ijk} \left\{ \alpha + (1-\alpha) \frac{A_{ik} S_{kj}}{X_{ij} \zeta_{ijk}} - \left(\frac{A_{ik} S_{kj}}{X_{ij} \zeta_{ijk}} \right)^{(1-\alpha)} \right\}, \tag{15}$$

with $\zeta_{ijk} = \frac{A_{ik}\widetilde{S}_{kj}}{\sum_l A_{il}\widetilde{S}_{lj}}$, *is an auxiliary function for*

$$F(\boldsymbol{S}) = \frac{1}{\alpha(1-\alpha)} \sum_{i,j} \alpha X_{ij} + (1-\alpha)[\boldsymbol{AS}]_{ij} - X_{ij}^{\alpha}[\boldsymbol{AS}]_{ij}^{1-\alpha}. \tag{16}$$

Proof. We need to show that the function $G(\boldsymbol{S}, \widetilde{\boldsymbol{S}})$ in (15) satisfies two conditions: (i) $G(\boldsymbol{S}, \boldsymbol{S}) = F(\boldsymbol{S})$; (ii) $G(\boldsymbol{S}, \widetilde{\boldsymbol{S}}) \geq F(\boldsymbol{S})$. One can easily see that $G(\boldsymbol{S}, \boldsymbol{S}) = F(\boldsymbol{S})$. Note that $\sum_k \zeta_{ijk} = 1$ from the definition and $\zeta_{ijk} \geq 0$ for $i = 1, \ldots, m$, $j = 1, \ldots, N$, and $k = 1, \ldots, n$. In order to prove that the condition (ii) is satisfied, we write $F(\boldsymbol{S})$ as

$$\begin{aligned} F(\boldsymbol{S}) =& \frac{1}{\alpha(1-\alpha)} \sum_{i,j} \alpha X_{ij} + (1-\alpha)[\boldsymbol{AS}]_{ij} - X_{ij}^{\alpha}[\boldsymbol{AS}]_{ij}^{1-\alpha} \\ =& \sum_{i,j} X_{ij} f\left(\frac{\sum_k A_{ik}S_{kj}}{X_{ij}}\right), \end{aligned} \tag{17}$$

where α-divergence is written using the convex function $f(\cdot)$ for positive α,

$$f(z) = \frac{1}{\alpha(1-\alpha)} \left\{\alpha + (1-\alpha)z - z^{1-\alpha}\right\}.$$

Jensen's inequality (due to the convexity of f) leads to

$$f\left(\sum_k A_{ik}S_{kj}\right) \leq \sum_k \zeta_{ijk} f\left(\frac{A_{ik}S_{kj}}{\zeta_{ijk}}\right). \tag{18}$$

Then, it follows from (18) that we have

$$F(\boldsymbol{S}) = \sum_{i,j} X_{ij} f\left(\frac{\sum_k A_{ik}S_{kj}}{X_{ij}}\right) \leq \sum_{i,j,k} X_{ij}\zeta_{ijk} f\left(\frac{A_{ik}S_{kj}}{X_{ij}\zeta_{ijk}}\right) = G(\boldsymbol{S}, \widetilde{\boldsymbol{S}}), \tag{19}$$

which proves the condition (ii). ■

Lemma 2 *Reversing the roles of* $\boldsymbol{S}$ *and* $\boldsymbol{A}$ *in Lemma 1, the function*

$$\begin{aligned} &G(\boldsymbol{A}, \widetilde{\boldsymbol{A}}) \\ &= \frac{1}{\alpha(1-\alpha)} \sum_{i,j,k} X_{ij}\xi_{ijk} \left\{\alpha + (1-\alpha)\frac{A_{ik}S_{kj}}{X_{ij}\xi_{ijk}} - \left(\frac{A_{ik}S_{kj}}{X_{ij}\xi_{ijk}}\right)^{(1-\alpha)}\right\}, \end{aligned} \tag{20}$$

with $\xi_{ijk} = \frac{\widetilde{A}_{ik}S_{kj}}{\sum_l \widetilde{A}_{il}S_{lj}}$, *is an auxiliary function for*

$$F(\boldsymbol{A}) = \frac{1}{\alpha(1-\alpha)} \sum_{i,j} \alpha X_{ij} + (1-\alpha)[\boldsymbol{AS}]_{ij} - X_{ij}^{\alpha}[\boldsymbol{AS}]_{ij}^{1-\alpha}. \tag{21}$$

Proof. This can be easily proved in the same way as Lemma 1. ■

Theorem 1 $D_\alpha[\boldsymbol{X}||\boldsymbol{AS}]$ *is non-increasing under the following multiplicative update rules:*

$$S_{ij} \leftarrow S_{ij} \left[\frac{\sum_k A_{ki} \left(X_{kj}/[\boldsymbol{AS}]_{kj} \right)^{\alpha}}{\sum_l A_{li}} \right]^{\frac{1}{\alpha}}, \tag{22}$$

$$A_{ij} \leftarrow A_{ij} \left[\frac{\sum_k S_{jk} \left(X_{ik}/[\boldsymbol{AS}]_{ik} \right)^{\alpha}}{\sum_l S_{jl}} \right]^{\frac{1}{\alpha}}. \tag{23}$$

Proof. The minimum of (15) is determined by setting the gradient to zero:

$$\frac{\partial G(\boldsymbol{S}, \widetilde{\boldsymbol{S}})}{\partial S_{ij}} = \frac{1}{\alpha} \sum_k A_{ki} \left\{ 1 - \left(\frac{A_{ki}S_{ij}}{X_{kj}\zeta_{kji}} \right)^{-\alpha} \right\} = 0, \tag{24}$$

which leads to

$$\left(\frac{S_{ij}}{\widetilde{S}_{ij}} \right)^{\alpha} = \left[\frac{\sum_k A_{ki} \left(\frac{X_{kj}}{\sum_l A_{kl}\widetilde{S}_{lj}} \right)^{\alpha}}{\sum_k A_{ki}} \right], \tag{25}$$

which suggest the updating rule for S_{ij}:

$$S_{ij} \leftarrow S_{ij} \left[\frac{\sum_k A_{ki} \left(X_{kj}/[\boldsymbol{AS}]_{kj} \right)^{\alpha}}{\sum_l A_{li}} \right]^{\frac{1}{\alpha}},$$

that is identical to (22).

In a similar manner, the updating rule (23) is determined by solving $\frac{\partial G(\boldsymbol{A}, \widetilde{\boldsymbol{A}})}{\partial \boldsymbol{A}} = 0$, where $G(\boldsymbol{A}, \widetilde{\boldsymbol{A}})$ is given in (20). ■

Multiplicative updates for our α-NMF are given in (22) and (23). These updates find a local minimum of $D_\alpha[\boldsymbol{X}||\boldsymbol{AS}]$. When $\alpha = 1$, (22) and (23) become equivalent to Lee-Seung NMF algorithm (Lee and Seung, 2001).

4 Alternative derivations and link with existing work

4.1 Projected gradient

The similar NMF algorithm in slightly different forms with over-relaxation and regularization terms has been proposed in (Cichocki et al., 2006b,a,c; Zdunek and Cichocki, 2006). However, our rigorous mathematical derivation and convergence analysis gives some new insight for this class of NMF algorithms. In this section, we show that updating rules (22) and (23) are also derived using the projected gradient (Cichocki et al., 2006a).

Partial derivatives of (14) with respect to $\boldsymbol{S}$ and $\boldsymbol{A}$, are given by

$$\frac{\partial D_\alpha[\boldsymbol{X}||\boldsymbol{AS}]}{\partial S_{ij}} = \frac{1}{\alpha}\left[\sum_k A_{ki} - \sum_k A_{ki}\left(\frac{X_{kj}}{[\boldsymbol{AS}]_{kj}}\right)^\alpha\right], \tag{26}$$

$$\frac{\partial D_\alpha[\boldsymbol{X}||\boldsymbol{AS}]}{\partial A_{ij}} = \frac{1}{\alpha}\left[\sum_k S_{jk} - \sum_k S_{jk}\left(\frac{X_{ik}}{[\boldsymbol{AS}]_{ik}}\right)^\alpha\right]. \tag{27}$$

The projected gradient method updates transformed parameters using the gradient information, which is of the form

$$\phi(S_{ij}) \leftarrow \phi(S_{ij}) - \eta_{ij}\frac{\partial D_\alpha[\boldsymbol{X}||\boldsymbol{AS}]}{\partial S_{ij}}, \tag{28}$$

$$\phi(A_{ij}) \leftarrow \phi(A_{ij}) - \eta_{ij}\frac{\partial D_\alpha[\boldsymbol{X}||\boldsymbol{AS}]}{\partial A_{ij}}, \tag{29}$$

where $\phi(\cdot)$ is a suitably-chosen function. Note that the the exponentiated gradient emerges for $\phi(\theta) = \log\theta$.

Thus, we have

$$S_{ij} \leftarrow \phi^{-1}\left(\phi(S_{ij}) - \eta_{ij}\frac{\partial D_\alpha[\boldsymbol{X}||\boldsymbol{AS}]}{\partial S_{ij}}\right), \tag{30}$$

$$A_{ij} \leftarrow \phi^{-1}\left(\phi(A_{ij}) - \eta_{ij}\frac{\partial D_\alpha[\boldsymbol{X}||\boldsymbol{AS}]}{\partial A_{ij}}\right). \tag{31}$$

Choosing $\phi(\theta) = \theta^\alpha$ and incorporating with (26) and (27), leads to (22) and (23).

4.2 KKT conditions

We show that the algorithms (22) and (23) can be also derived using the KKT conditions. The minimization of $D_\alpha[\boldsymbol{X}||\boldsymbol{AS}]$ in (14) with nonnegativity constraints, $A_{ij} \geq 0$, $S_{ij} \geq 0$, can be formulated as a constrained minimization problem with inequality constraints. Denote by $\Lambda_{ij} \geq 0$ and $\Omega_{ij} \geq 0$ Lagrangian multipliers associated with constraints, $A_{ij} \geq 0$, $S_{ij} \geq 0$, respectively.

The KKT conditions require:

$$\frac{\partial D_\alpha[\boldsymbol{X}||\boldsymbol{AS}]}{\partial S_{ij}} = \Omega_{ij}, \tag{32}$$

$$\frac{\partial D_\alpha[\boldsymbol{X}||\boldsymbol{AS}]}{\partial A_{ij}} = \Lambda_{ij}, \tag{33}$$

as optimality conditions and

$$\Omega_{ij} S_{ij} = 0, \tag{34}$$

$$\Lambda_{ij} A_{ij} = 0, \tag{35}$$

as complementary slackness conditions, implying that

$$\Omega_{ij} S_{ij}^\alpha = 0, \tag{36}$$

$$\Lambda_{ij} A_{ij}^\alpha = 0. \tag{37}$$

Multiplying both sides of (32) and (33) by S_{ij}^α and A_{ij}^α, respectively, and incorporating with (36) and (37), leads to

$$\frac{1}{\alpha}\left[\sum_k A_{ki} - \sum_k A_{ki}\left(\frac{X_{kj}}{[\boldsymbol{AS}]_{kj}}\right)^\alpha\right] S_{ij}^\alpha = 0, \tag{38}$$

$$\frac{1}{\alpha}\left[\sum_k S_{jk} - \sum_k S_{jk}\left(\frac{X_{ik}}{[\boldsymbol{AS}]_{ik}}\right)^\alpha\right] A_{ij}^\alpha = 0, \tag{39}$$

which suggests iterative algorithms

$$S_{ij} \leftarrow S_{ij}\left[\frac{\sum_k A_{ki}\left(X_{kj}/[\boldsymbol{AS}]_{kj}\right)^\alpha}{\sum_l A_{li}}\right]^{\frac{1}{\alpha}}, \tag{40}$$

$$A_{ij} \leftarrow A_{ij}\left[\frac{\sum_k S_{jk}\left(X_{ik}/[\boldsymbol{AS}]_{ik}\right)^\alpha}{\sum_l S_{jl}}\right]^{\frac{1}{\alpha}}. \tag{41}$$

5 Numerical experiments

5.1 Image denoising

We apply our NMF algorithm to the task of image denosing, in which reconstructed images using learned basis images and encoding variables are associated with denoised imaged. We consider ORL face DB (Samaria and Harter, 1994) (40 people and 10 images for each person, i.e., $40 \times 10 = 400$ images in total). 3 different types of noise are considered, For each image, 5% of pixels are randomly chosen and then are converted to black or white pixels, producing pepper (black), salt (white), and pepper & salt (black and white) images (see Fig .1).

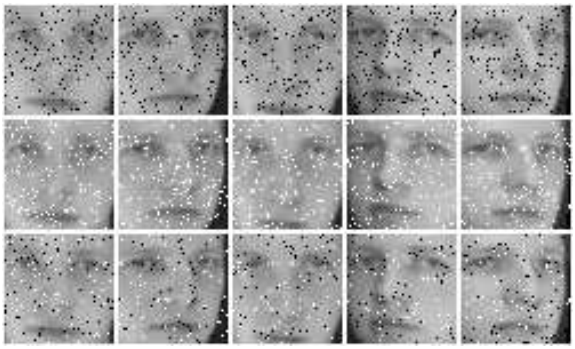
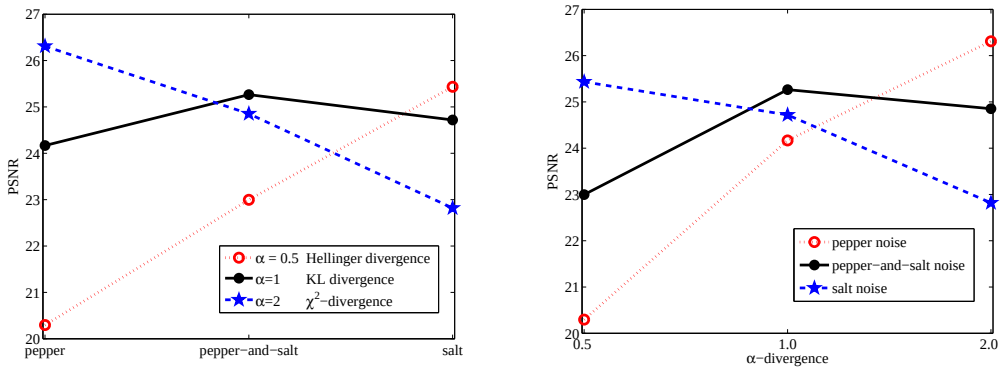


Fig. 1. From top to bottom: images contaminated by pepper (black), salt (white), and pepper & salt (black and white) noise.

We apply our NMF algorithm with different values of α ($\alpha = 0.5, 1, 2$). Experiments are carried out 30 times independently for each type of noise and each value of $\alpha = \{0.5, 1, 2\}$. As a performance measure, averaged peak-signal-to-noise ratio (PSNR) is used. Higher PSNR values represent better results. Fig .2 shows some interesting behavior. The larger α results in the better performance in the case of pepper noise and the smaller α works better in the case of salt noise. In fact, these results are consistent with the characteristics of α-divergence where $D_\alpha[p||q]$ emphasizes the part where p is small as α increases Amari (2007).

5.2 EEG classification

We apply our α-divergence-based NMF algorithm (multiplicative updates are described in Eqs. (22) and (23)) to a problem of EEG classification and evaluate the performance for several different values of α ($\alpha = 0.5, 1, 2$). EEG classification plays a very important role in brain computer interface (BCI) where a subject's mental state is required to be estimated from EEG signals.



(a) (b)

Fig. 2. The relationship between types of noise and divergences. In the case of pepper (black) noise, higher *alpha* value gives robust result. In the case of salt (white) noise, lower *alpha* values gives robust result.

Our previous work (Lee et al., 2006) demonstrated that NMF could extract spectral features that are useful in EEG classification. Two EEG data sets that were used in our empirical study are as follows:

- **Graz dataset:** Dataset III in BCI competition II, which was provided by the Laboratory of Brain-Computer Interfaces (BCI-Lab), Graz University of Technology (Blankertz et al., 2004; Lemm et al., 2004);
- **IDIAP dataset:** Dataset V in BCI competition III, which was provided by the IDIAP Research Institute (J. del R. Millán, 2004).

5.2.1 Graz dataset

Graz dataset involves left/right imagery hand movements and consists of 140 labeled trials for training and 140 unlabeled trials for test. Each trial has a duration of 9 seconds, where a visual cue (arrow) is presented pointing to the left or the right after 3-second preparation period and the imagination of left or right movement is carried out for 6 seconds. It contains EEG acquired from three different channels (with sampling frequency 128 Hz) C_3, C_z and C_4. In our experiment we use only two channels, C_3 and C_4, because C_z channel contains little information for discriminant analysis. Requirements for result comparison is to provide a continuous classification accuracy for each time point of trial during imagination session.

Time-domain EEG signals are converted into time-frequency representations by complex Morlet wavelet transform. Then we apply the NMF algorithm with $\alpha = 0.5, 1, 2$ and $n = 2, 4, 5, 6$ (the number of basis vectors), in order to estimate basis vectors shown in Fig. 3. As the number of basis vector increases, the spectral components such as μ rhythm (8-12 Hz), β rhythm (18-22 Hz), and sensori-motor rhythm (12-16 Hz) appear in the order of their importance.

All rhythms have the property of contralateral dominance, so they are present in basis vectors associated with C_3 or C_4 channel, separately.

Feature vectors correspond to the column vectors of the encoding variable matrix $\boldsymbol{S}$. We use the same probabilistic model-based classifier as used in (Lemm et al., 2004; Lee et al., 2006). The best performance in this experiment was achieved when $\alpha = 0.5$or1 and $n = 5$, as summarized in Table 1. The maximal classification accuracy is 88.57 % at 6.05 sec the mutual information (MI) hits the maximum, 0.6549 bit, which occurs at 6.05 sec. The result is better than the one achieved by the BCI competition 2003 winner (0.61 bit). Table 1 shows the maximum mutual information in the time courses per trial with α and n varying. The smaller the value of α, the better the mutual information, however, α is not a critical factor in this experiment, since NMF works pretty well across different values of α.

Table 1
Mutual information between the true class label and the estimated class label in cases of different values of α and n.

	number of basis vectors				
	$n = 2$	$n = 4$	$n = 5$	$n = 6$	$n = 7$
$\alpha = 0.5$	0.5545	0.5803	**0.6549**	0.6256	0.5875
$\alpha = 1$	0.5545	0.5803	**0.6549**	0.6256	0.5803
$\alpha = 2$	0.5408	0.5745	0.6404	0.6256	0.5803

5.2.2 IDIAP dataset

IDIAP dataset contains EEG data recorded from 3 normal subjects during 4 non-feedback sessions, involving three tasks which include the imagination of repetitive self-paced left/right hand movements and the generation of words beginning with the same random letter. All 4 sessions were acquired on the same day, each lasting 4 minutes with 5-10 minutes breaks in between them. The subject performed a given task for about 15 seconds and then switched randomly to another task at the operator's request. In contrast to the Graz dataset, EEG data is not split into trials (i.e., no trial structure), since subjects are continuously performing any of mental tasks.

Data are provided in two ways: (1) raw EEG signals (with sampling rate = 512 Hz) recorded from 32 electrodes; (2) precomputed features. We use the precomputed features in experiments. They were obtained by the power spectral density (PSD) in the frequency band 8-30 Hz every 62.5 ms, (i.e., 16 times per second) over the last second of data with a frequency resolution of 2 Hz for the eight centro-parietal channels C_3, C_z, C_4, CP_1, CP_2, P_3, P_z, and P_4 after the raw EEG potentials were first spatially filtered by means of a surface

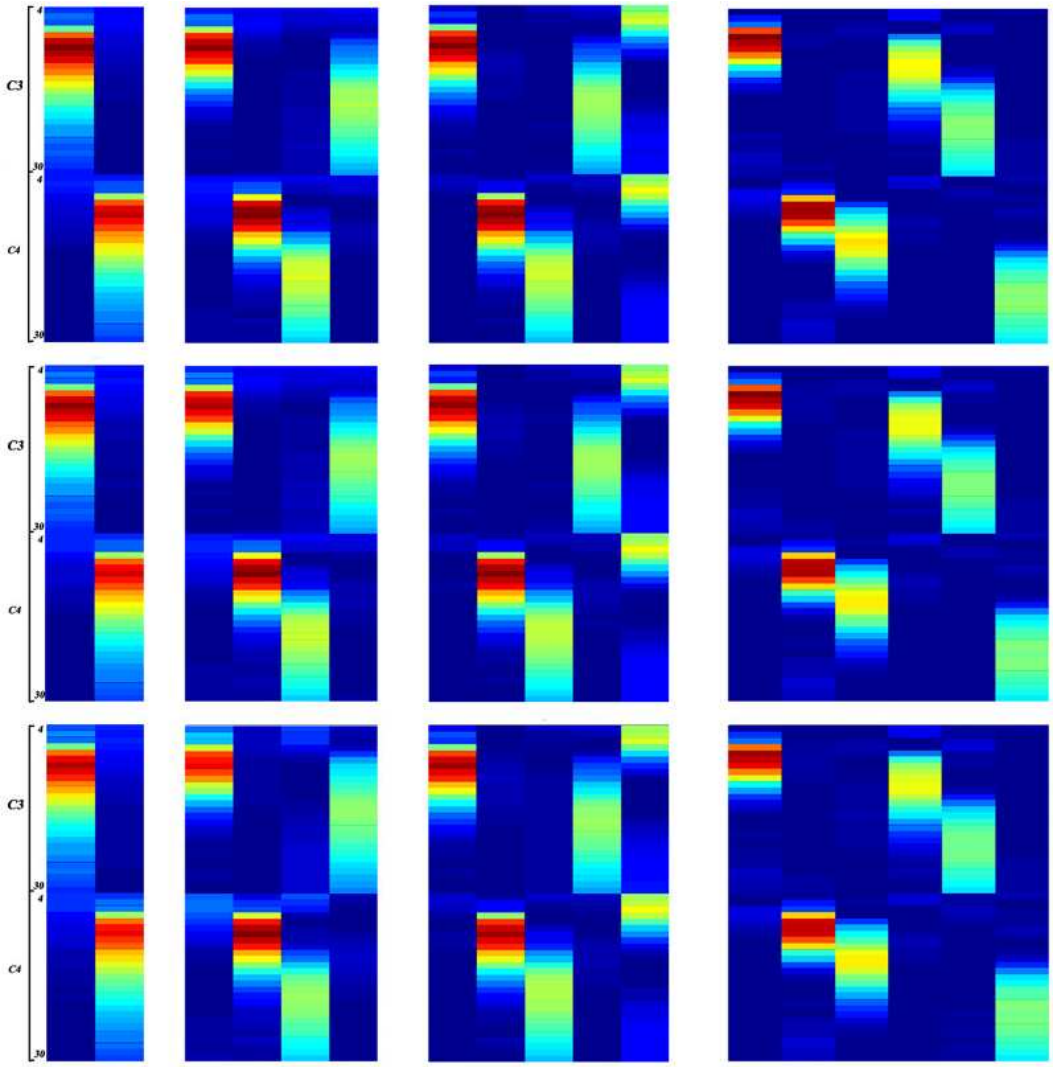


Fig. 3. Basis vectors determined by NMF are shown in the case of $\alpha = 0.5, 1, 2$ (from top to bottom) and $n = 2, 4, 5, 6$ (from left to right). In each plot, top 1/2 is associated with C_3 and bottom 1/2 is contributed by C_4. In each of those, the vertical axis represents frequencies between 4 and 30 Hz, the horizon axis is related to the number of basis vectors. For example, the upper-left one ($\alpha = 0.5$ and $n = 2$) is associated with 54×2, where the upper half corresponds to 27 frequency bins over[4,30] Hz for C_3 (the first row corresponds to the power associated with 4 Hz) and the lower half is also associated with 27 frequency bins for C_4. Basis vectors reveals some useful characteristics: (1) μ rhythm (8-12 Hz); (2) β rhythm (18-22 Hz); (3) sensori-motor rhythm (12-16 Hz). ERD has the contralateral dominance, hence each rhythm occurs in each channel separately. Different values of α provide slightly different basis vectors, although the distinction is very small since all of those basis vectors well represent discriminative spectral characteristics.

Laplacian. As a result, an EEG sample is a 96-dimensional vector (eight channels times 12 frequency components). Requirements for comparative study are to provide an output every 0.5 second using the last second of data.

We use the precomputed features, then we don't need any preprocessing procedure except for normalization of each data vector. We apply the NMF algorithm with $\alpha = 0.5, 1, 2$ and $n = 3, \ldots, 9$ (the number of basis vectors). For the on-line classification for IDIAP data which consist of uncued EEG signals, we use the Viterbi algorithm (Forney, 1973) that is a dynamic programming algorithm for finding a most probable sequence of hidden states that explains a sequence of observations.

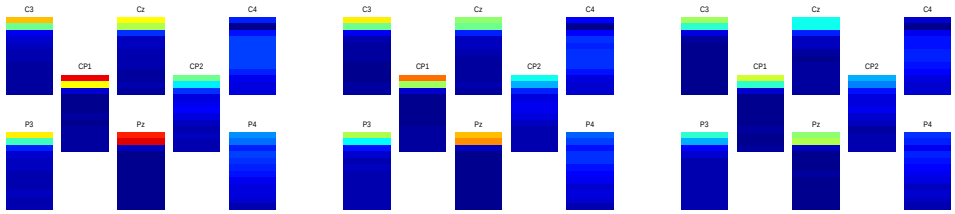
Fig. 4 shows basis vectors computed by NMF for $\alpha = 0.5, 1, 2$. Spectral characteristics of basis vectors is shown in Fig. 5, where $\alpha = 0.5$ reveals the most discriminative characteristics (sharper peaks in spectrum compared to cases of $\alpha = 1$ and $\alpha = 2$). Table 2 summarizes the classification result in the case of $n = 4$, where the result of BCI competition winner is also included. NMF with $\alpha = 0.5$ gives the best performance in this experiment. Basis vectors also show strong activations in μ rhythm band as well as lower alpha band, as shown in Figs. 4 and 5. In general, it is an open problem how to select the optimal value of α, since it varies depending on datasets.

Table 2
Classification accuracy for IDIAP dataset.

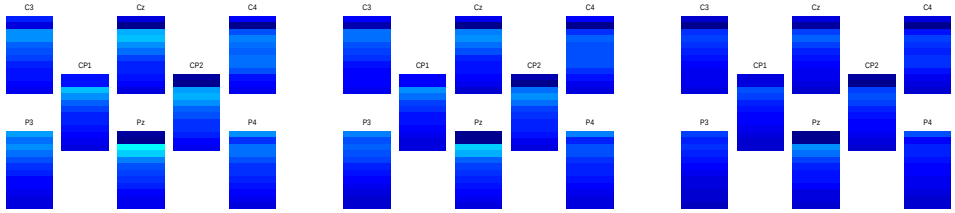
	sub1	sub2	sub3	avg
$\alpha = 0.5$	**84.93**	**77.19**	**58.49**	**73.54**
$\alpha = 1$	83.56	70.97	57.11	70.55
$\alpha = 2$	81.51	74.19	52.29	69.33
BCI comp. winner	79.60	70.31	56.02	68.65

6 Conclusions

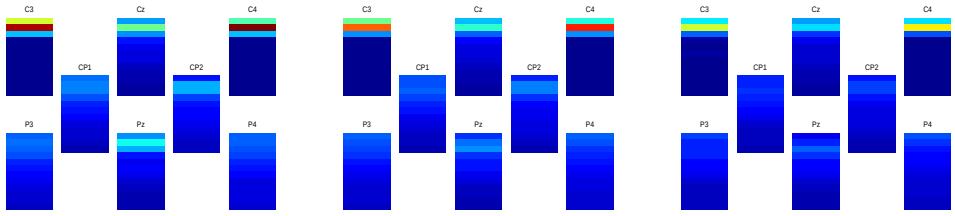
We have presented multiplicative updates for NMF which iteratively minimize the Amari's α-divergence between the observed data and the model. Our multiplicative updates include some existing NMF algorithm as their special cases, since α-divergence is a parametric divergence measure which contains KL-divergence, Hellinger divergence, χ^2-divergence, and so on. The hyperparameter α is associated with the characteristics of a learning machine, in the sense that the model distribution is more inclusive (as α goes to ∞) more exclusive (as α approaches $-\infty$). We have also presented alternative derivation of the algorithm using the projected gradient and KKT conditions. The paper has focused primarily on the derivation and monotonic convergence of a generic α-divergence-based NMF algorithm. We have applied our method to



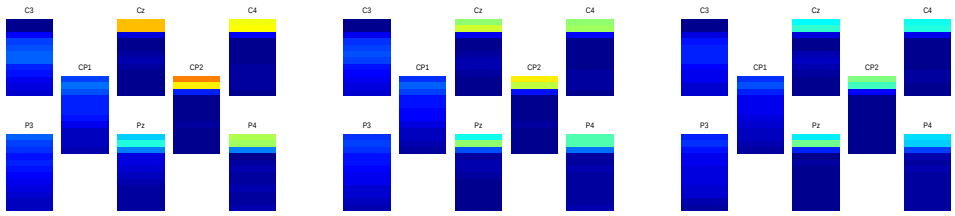
(a) 1st basis, $\alpha = 0.5$ (b) 1st basis, $\alpha = 1$ (c) 1st basis, $\alpha = 2$



(d) 2nd basis, $\alpha = 0.5$ (e) 2nd basis, $\alpha = 1$ (f) 2nd basis, $\alpha = 2$



(g) 3rd basis, $\alpha = 0.5$ (h) 3rd basis, $\alpha = 1$ (i) 3rd basis, $\alpha = 2$



(j) 4th basis, $\alpha = 0.5$ (k) 4th basis, $\alpha = 1$ (l) 4th basis, $\alpha = 2$

Fig. 4. Basis vectors determined by NMF are shown in the case of $\alpha = 0.5, 1, 2$ (from left to right) and n=4 (from top to down). In each plot, basis vectors are associated with the power spectrum for 8 channels, C_3, C_Z, C_4, CP_1, CP_2, P_3, P_Z, and P_4, while the vertical axis in each basis vector corresponds to frequency bins $8, 10, 12, \ldots, 28, 30Hz$. Three different tasks involve the imagination of left/right hand movements and word generation. The first and last basis vectors show activations in the frequency band $8 - 12Hz$ (μ rhythm), either in left or right hemisphere channels, indicating that left or right movements are involved. In general, the EEG phenomenon related to word generation is not well known, compared to motor imagery task. We guess that the word generation is related to the lower alpha band $(8 - 10Hz)$.

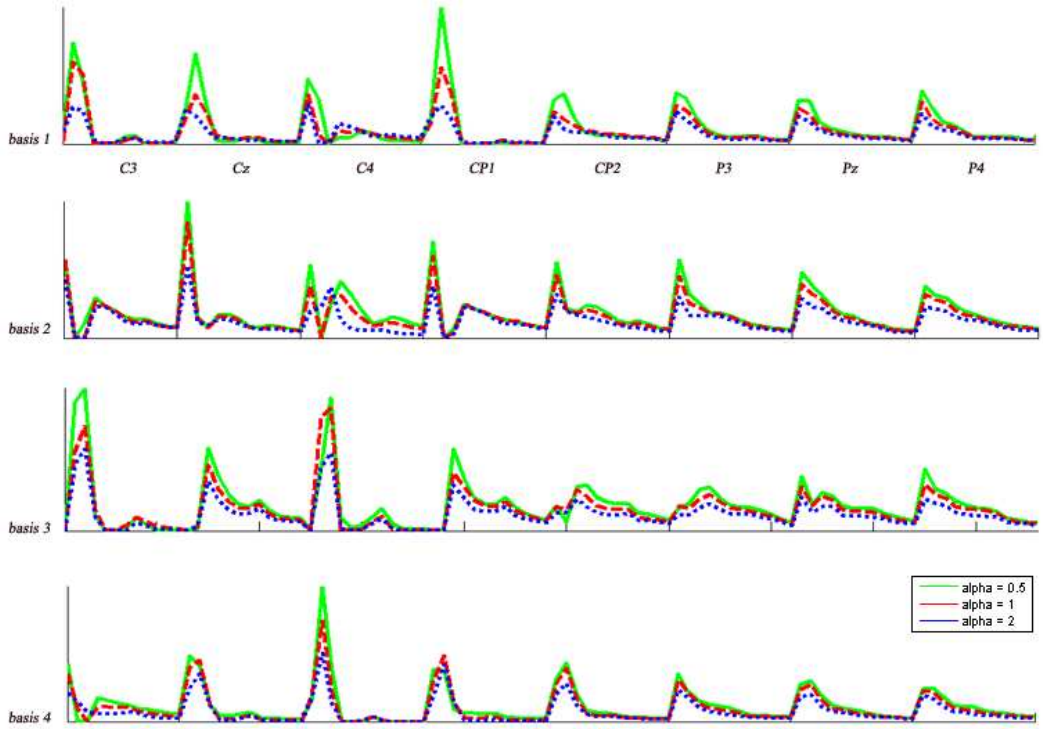


Fig. 5. Spectral characteristics of basis vectors computed by NMF for $\alpha = 0.5, 1, 2$. Each unit in Fig. 4 is a 12-dimensional vectors. In NMF computation we concatenate 8 channels into a single vector, leading to 96-dimensional basis vectors. The plot shows the power spectrum for each basis vector, with respect to $[8, 10, 12, \ldots, 28, 30]Hz$ in the horizontal axis. In the case of $\alpha = 0.5$, the strongest peaks are observed, compared to other cases.

the task of image denoising and feature extraction for EEG classification. In the case of image denoising, we have observed that the larger α produced the better performance in the case of pepper noise and the smaller α worked better in the case of salt noise. In fact, these results are consistent with the characteristics of α-divergence where $D_\alpha[p||q]$ emphasizes the part where p is small as α increases Amari (2007). We have also investigated the EEG classification performance for several different values of α. Different values of α did not have much influence on determining basis vectors in the case of a well-separated task such as motor imagery (Graz dataset in Experiment 1). However, in the tasks such as motor imagery mixed with word generation (IDIAP dataset in Experiment 2), it was observed that α played a critical role in determining discriminative basis vectors. In general, it is still an open problem how to select an appropriate value of α, since it varies across datasets. The main contribution of this paper is in the derivation of generic multiplicative updates for NMF and its convergence study when α-divergence is used as a discrepancy measure in the context of NMF.

Acknowledgments: Portion of this work was carried out when S. Choi was

visiting Laboratory for Advanced Brain Signal Processing, RIKEN BSI, Japan. This work was supported by National Core Research Center for Systems Bio-Dynamics and KOSEF Basic Research Program (grant R01-2006-000-11142-0).

References

Ahn, J. H., Kim, S., Oh, J. H., Choi, S., 2004. Multiple nonnegative-matrix factorization of dynamic PET images. In: Proceedings of Asian Conference on Computer Vision.

Ali, S. M., Silvey, S. D., 1966. A general class of coefficients of divergence of one distribution from another. Journal of the Royal Statistical Society B 28, 131–142.

Amari, S., 1985. Differential Geometrical Methods in Statistics. Springer.

Amari, S., 2007. Integration of stochastic models by minimizing α-divergence. Neural Computation 19 (10), 2780–2796.

Blankertz, B., Müller, K. R., Curio, G., Vaughan, T. M., Schalk, G., Wolpaw, J. R., Schlögl, A., Neuper, C., Pfurtscheller, G., Hinterberger, T., Schroder, M., Birbaumer, N., 2004. The BCI competition 2003: Progress and perspectives in detection and discrimination of EEG single trials. IEEE Trans. Biomedical Engineering 51 (6).

Brunet, J. P., Tamayo, P., Golub, T. R., Mesirov, J. P., 2004. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences, USA 101 (12), 4164–4169.

Cho, Y. C., Choi, S., 2005. Nonnegative features of spectro-temporal sounds for classfication. Pattern Recognition Letters 26 (9), 1327–1336.

Cichocki, A., Amari, S., Zdunek, R., 2006a. Extended SMART algorithms for non-negative matrix factorization. In: Proceedings of the Eighth International Conference on Artificial Intelligence and Soft Computing. Zakopane, Poland.

Cichocki, A., Zdunek, R., Amari, S., 2006b. Csiszár's divergences for non-negative matrix factorization: Family of new algorithms. In: Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation. Charleston, South Carolina.

Cichocki, A., Zdunek, R., Amari, S., 2006c. New algorithms for non-negative matrix factorization in applications to blind source separation. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Toulouse, France.

Csiszár, I., 1974. Information measures: A critical survey. In: Trans. 7th Prague Conference on Information Theory. Vol. A. pp. 73–86.

Dhillon, I. S., Sra, S., 2006. Generalized nonnegative matrix approximations with Bregman divergences. In: Advances in Neural Information Processing Systems. Vol. 18. MIT Press.

Forney, G. D., 1973. The Viterbi algorithm. Proceedings of of the IEEE 61, 268–278.

Hoyer, P. O., 2004. Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research 5, 1457–1469.

J. del R. Millán, 2004. On the need for on-line learning in brain-computer interfaces. In: Proceedings of the International Joint Conference on Neural Networks. Budapest, Hungary.

Kim, M., Choi, S., 2006. Monaural music source separation: Nonnegativity, sparseness, and shift-invariance. In: Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation. Springer, Charleston, South Carolina, pp. 617–624.

Kompass, R., 2007. A generalized divergence measure for nonnegative matrix factorization. Neural Computation 19, 780–791.

Lee, D. D., Seung, H. S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791.

Lee, D. D., Seung, H. S., 2001. Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems. Vol. 13. MIT Press.

Lee, H., Cichocki, A., Choi, S., 2006. Nonnegative matrix factorization for motor imagery EEG classification. In: Proceedings of the International Conference on Artificial Neural Networks. Springer, Athens, Greece.

Lee, J. S., Lee, D. D., Choi, S., Lee, D. S., 2001. Application of non-negative matrix factorization to dynamic positron emission tomography. In: Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation. San Diego, California, pp. 629–632.

Lemm, S., Schäfer, C., Curio, G., 2004. BCI competition 2003-data set III: Probabilistic modeling of sensorimotor μ rhythms for classification of imaginary hand movements. IEEE Trans. Biomedical Engineering 51 (6).

Li, S. Z., Hou, X. W., Zhang, H. J., Cheng, Q. S., 2001. Learning spatially localized parts-based representation. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Kauai, Hawaii, pp. 207–212.

Lin, C. J., 2007. On the convergence of multiplicative update algorithms for non-negative matrix factorization. IEEE Trans. Neural Networks (to appear).

Paatero, P., Tapper, U., 1997. Least squares formulation of robust non-negative factor analysis. Chemometrics Intelligent Laboratory Systems 37, 23–35.

Samaria, F., Harter, A., 1994. Parameterisation of a stochastic model for human face identification. In: Proceedings of the 2nd IEEE Workshop on Applications of Computer Vision. Sarasota, FL.

Shahnaz, F., Berry, M., Pauca, P., Plemmons, R., 2006. Document clustering using nonnegative matrix factorization. Information Processing and Management 42, 373–386.

Smaragdis, P., Brown, J. C., 2003. Non-negative matrix factorization for poly-

phonic music transcription. In: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz, NY, pp. 177–180.

Xu, W., Liu, X., Gong, Y., 2003. Document clustering based on non-negative matrix factorization. In: Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval. Toronto, Canada.

Zdunek, R., Cichocki, A., 2006. Non-negative matrix factorization with quasi-Newton optimization. In: Proceedings of the Eighth International Conference on Artificial Intelligence and Soft Computing. Zakopane, Poland.

Zhu, H., Rohwer, R., 1995. Information geometric measurements of generalization. Tech. Rep. NCRG-4350, Neural Computing Research Group, Aston University.