Entropy **2011**, *13*, 134-170; doi:10.3390/e13010134

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization

Andrzej Cichocki [1,2,⋆], Sergio Cruces [3,⋆] and Shun-ichi Amari [4]

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

[2] Systems Research Institute, Intelligent Systems Laboratory, PAS, Newelska 6 str., 01-447 Warsaw, Poland

[3] Dpto de Teoría de la Señal y Comunicaciones, University of Seville, Camino de los Descubrimientos s/n, 41092-Seville, Spain

[4] Laboratory for Mathematical Neuroscience, RIKEN BSI, Wako, 351-0198 Saitama, Japan; E-Mail: amari@brain.riken.jp

⋆ Authors to whom correspondence should be addressed; E-Mails: a.cichocki@riken.jp (A.C.); sergio@us.es (S.C.).

Received: 13 December 2010; in revised form: 4 January 2011 / Accepted: 4 January 2011 / Published: 14 January 2011

Abstract: We propose a class of multiplicative algorithms for Nonnegative Matrix Factorization (NMF) which are robust with respect to noise and outliers. To achieve this, we formulate a new family generalized divergences referred to as the Alpha-Beta-divergences (AB-divergences), which are parameterized by the two tuning parameters, alpha and beta, and smoothly connect the fundamental Alpha-, Beta- and Gamma-divergences. By adjusting these tuning parameters, we show that a wide range of standard and new divergences can be obtained. The corresponding learning algorithms for NMF are shown to integrate and generalize many existing ones, including the Lee-Seung, ISRA (Image Space Reconstruction Algorithm), EMML (Expectation Maximization Maximum Likelihood), Alpha-NMF, and Beta-NMF. Owing to more degrees of freedom in tuning the parameters, the proposed family of AB-multiplicative NMF algorithms is shown to improve robustness with respect to noise and outliers. The analysis illuminates the links of between AB-divergence and other divergences, especially Gamma- and Itakura-Saito divergences.

Keywords: nonnegative matrix factorization (NMF); robust multiplicative NMF algorithms; similarity measures; generalized divergences; Alpha-; Beta-; Gamma- divergences; extended Itakura-Saito like divergences; generalized Kullback-Leibler divergence

1. Introduction

In many applications such as image analysis, pattern recognition and statistical machine learning, it is beneficial to use the information-theoretic divergences rather than the Euclidean squared distance. Among them the Kullback-Leibler, Hellinger, Jensen-Shannon and Alpha-divergences have been pivotal in estimating similarity between probability distributions [1–6]. Such divergences have been successfully applied as cost functions to derive multiplicative and additive projected gradient algorithms for nonnegative matrix and tensor factorizations [7–10]. Such measures also play an important role in the areas of neural computation, pattern recognition, estimation, inference and optimization. For instance, many machine learning algorithms for classification and clustering employ a variety of (dis)similarity measures which are formulated using information theory, convex optimization, and information geometry [11–27]. Apart from the most popular squared Euclidean distance and Kullback-Leibler divergence, recently, alternative generalized divergences such as the Csiszár-Morimoto f-divergence and Bregman divergences have become attractive alternatives for advanced machine learning algorithms [7–10,28–34].

In this paper, we present a novel (dis)similarity measure which smoothly connects or integrates many existing divergences and allows us to develop new flexible and robust algorithms for NMF and related problems. The scope of the results presented in this paper is vast, since the generalized Alpha-Beta-divergence (or simply AB-divergence) function includes a large number of useful cost functions containing those based on the relative entropies, generalized Kullback-Leibler or I-divergence, Hellinger distance, Jensen-Shannon divergence, J-divergence, Pearson and Neyman Chi-square divergences, Triangular Discrimination and Arithmetic-Geometric divergence. Furthermore, the AB-divergence provides a natural extension of the families of Alpha- and Beta-divergences, gives smooth connections between them and links to other fundamental divergences. The divergence functions discussed in this paper are flexible because they allow us to generate a large number of well-known and often used particular divergences (for specific values of tuning parameters). Moreover, by adjusting adaptive tuning parameters of factor matrices, we can optimize the cost function for NMF algorithms and estimate desired parameters of the underlying factorization model in the presence of noise and/or outliers.

1.1. Introduction to NMF and Basic Multiplicative Algorithms for NMF

The Nonnegative Matrix Factorization (NMF) problem has been investigated by many researchers, e.g., Paatero and Tapper [36], and has gained popularity through the work of Lee and Seung [37,38]. Based on the argument that the nonnegativity is important in human perception, they proposed simple algorithms (often called the Lee-Seung algorithms) for finding physically meaningful nonnegative representations of nonnegative signals and images [38].

The basic NMF problem can be stated as follows: Given a nonnegative data matrix $\mathbf{Y} = \mathbf{P} = [p_{it}] \in \mathbb{R}_+^{I\times T}$ (with $p_{it} \geq 0$ or equivalently $\mathbf{P} \geq \mathbf{0}$) and a reduced rank J ($J \leq \min(I,T)$, typically $J << \min(I,T)$), find two nonnegative matrices $\mathbf{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_J] \in \mathbb{R}_+^{I\times J}$ and $\mathbf{X} = \mathbf{B}^T = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_J]^T \in \mathbb{R}_+^{J\times T}$ which factorize $\mathbf{P}$ as faithfully as possible, that is

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E} = \mathbf{A}\mathbf{B}^T + \mathbf{E}, \tag{1}$$

where the matrix $\mathbf{E} \in \mathbb{R}^{I\times T}$ represents approximation error. Since we usually operate on column vectors of matrices for convenience we shall use often the matrix $\mathbf{B} = \mathbf{X}^T$ instead of the matrix $\mathbf{X}$.

The factors $\mathbf{A}$ and $\mathbf{X}$ may have different physical meanings in different applications: in a Blind Source Separation (BSS) $\mathbf{A}$ denotes mixing matrix and $\mathbf{X}$ source signals; in clustering problems, $\mathbf{A}$ is the basis matrix and $\mathbf{X}$ is a weight matrix; in acoustic analysis, $\mathbf{A}$ represents the basis patterns, while each row of $\mathbf{X}$ corresponds to sound patterns activation [7].

Standard NMF only assumes nonnegativity of factor matrices $\mathbf{A}$ and $\mathbf{X}$. Unlike blind source separation methods based on independent component analysis (ICA), here we do not assume that the sources are independent, although we can impose some additional constraints such as smoothness, sparseness or orthogonality of $\mathbf{A}$ and/or $\mathbf{X}$ [7].

Although the NMF can be applied to the BSS problems for nonnegative sources and nonnegative mixing matrices, its application is not limited to the BSS and it can be used in various and diverse applications far beyond BSS [7].

In NMF, our aim is to find the entries of nonnegative matrices $\mathbf{A}$ and $\mathbf{X}$ assuming that a data matrix $\mathbf{P}$ is known:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E} = \mathbf{A}\mathbf{B}^T + \mathbf{E} = \sum_{j=1}^{J} \boldsymbol{a}_j \boldsymbol{b}_j^T + \mathbf{E}, \tag{2}$$

or equivalently in a scalar form:

$$y_{it} = p_{it} = \sum_{j=1}^{J} a_{ij} x_{jt} + e_{it} = \sum_{j=1}^{J} a_{ij} b_{tj} + e_{it}. \tag{3}$$

In order to estimate nonnegative factor matrices $\mathbf{A}$ and $\mathbf{X}$ in the standard NMF, we need to consider the similarity measures to quantify a difference between the data matrix $\mathbf{P}$ and the approximative NMF model matrix $\widehat{\mathbf{P}} = \mathbf{Q} = \mathbf{A}\mathbf{X}$.

The choice of the similarity measure (also referred to as distance, divergence or measure of dissimilarity) mostly depends on the probability distribution of the estimated signals or components and on the structure of data and a distribution of a noise.

The best known and the most frequently used adaptive multiplicative algorithms for NMF are based on the two loss functions: Squared Euclidean distance and generalized Kullback-Leibler divergence also called the I-divergence.

The squared Euclidean distance is based on the Frobenius norm:

$$D_E(\mathbf{Y}||\mathbf{A}\mathbf{X}) = \frac{1}{2}||\mathbf{Y} - \mathbf{A}\mathbf{X}||_F^2, \tag{4}$$

which is optimal for additive Gaussian noise [7]. It should be noted that the above cost function is convex with respect to either elements of matrix $\mathbf{A}$ or matrix $\mathbf{X}$, but not both.

Remark: Although the NMF optimization problem is not convex, the objective functions are separately convex in each of the two factors $\mathbf{A}$ and $\mathbf{X}$, which implies that finding the optimal factor matrix $\mathbf{A}$ corresponding to a fixed matrix $\mathbf{X}$ reduces to a convex optimization problem and *vice versa*. However, the convexity is lost as soon as we try to optimize factor matrices simultaneously [39].

Using a gradient descent approach for cost function (4) and switching alternatively between the two sets of parameters, we obtain the simple multiplicative update formulas (see Section 3 for derivation of algorithms in a more general form):

$$a_{ij} \leftarrow a_{ij} \frac{[\mathbf{P}\,\mathbf{X}^T]_{ij}}{[\mathbf{Q}\,\mathbf{X}^T]_{ij} + \varepsilon}, \tag{5}$$

$$x_{jt} \leftarrow x_{jt} \frac{[\mathbf{A}^T\,\mathbf{P}]_{jt}}{[\mathbf{A}^T\mathbf{Q}]_{jt} + \varepsilon}, \tag{6}$$

where a small positive constant ε prevents division by zero, $\mathbf{P} = \mathbf{Y}$ and $\mathbf{Q} = \mathbf{A}\mathbf{X}$.

The above algorithm (5)-(6), called often the Lee-Seung NMF algorithm can be considered as a natural extension of the well known algorithm ISRA proposed first by Daube-Witherspoon and Muehllehner [40] and investigated extensively by De Pierro and Byrne [41–45].

The above update rules can be written in a compact matrix form as

$$\mathbf{A} \leftarrow \mathbf{A} \circledast \left[(\mathbf{P}\mathbf{X}^T) \oslash (\mathbf{Q}\mathbf{X}^T + \varepsilon) \right], \tag{7}$$

$$\mathbf{X} \leftarrow \mathbf{X} \circledast \left[(\mathbf{A}^T\mathbf{P}) \oslash (\mathbf{A}^T\mathbf{Q} + \varepsilon) \right], \tag{8}$$

where $\circledast$ is the Hadamard (components-wise) product and $\oslash$ is element-wise division between two matrices. In practice, the columns of the matrix $\mathbf{A}$ should be normalized to the unit ℓ_p-norm (typically, $p = 1$).

The original ISRA algorithm is relatively slow, and several heuristic approaches have been proposed to speed it up. For example, a relaxation approach rises the multiplicative coefficients to some power $w \in (0, 2]$, that is [7,47],

$$a_{ij} \leftarrow a_{ij} \left(\frac{[\mathbf{P}\,\mathbf{X}^T]_{ij}}{[\mathbf{Q}\,\mathbf{X}^T]_{ij}} \right)^w, \tag{9}$$

$$x_{jt} \leftarrow x_{jt} \left(\frac{[\mathbf{A}^T\,\mathbf{P}]_{jt}}{[\mathbf{A}^T\mathbf{Q}]_{jt}} \right)^w, \tag{10}$$

in order to achieve faster convergence.

Another frequently used cost function for the NMF is the generalized Kullback-Leibler divergence (also called the I-divergence) [38]:

$$D_{KL}(\mathbf{P}||\mathbf{Q}) = \sum_{it} \left(p_{it} \, \ln \frac{p_{it}}{q_{it}} - p_{it} + q_{it} \right), \tag{11}$$

where $\mathbf{Q} = \hat{\mathbf{P}} = \mathbf{A}\mathbf{X}$ with entries $q_{it} = [\mathbf{A}\mathbf{X}]_{it}$.

Similar to the squared Euclidean cost function, the I-divergence is convex with respect to either $\mathbf{A}$ or $\mathbf{X}$, but it is not generally convex with respect to $\mathbf{A}$ and $\mathbf{X}$ jointly, so the minimization of such a cost function can yield many local minima.

By minimizing the cost function (11) subject to the nonnegativity constraints, we can easily derive the following multiplicative learning rule, referred to as the EMML algorithm (Expectation Maximization Maximum Likelihood) [44,45]. The EMML algorithm is sometimes called the Richardson-Lucy algorithm (RLA) or simply the EM algorithm [48–50]. In fact, the EMML algorithm was developed for a fixed and known $\mathbf{A}$. The Lee-Seung algorithm (which is in fact the EMML algorithm) based on the I-divergence employs alternative switching between $\mathbf{A}$ and $\mathbf{X}$ [37,38]:

$$x_{jt} \leftarrow x_{jt} \frac{\sum_{i=1}^{I} a_{ij} \left(p_{it}/q_{it}\right)}{\sum_{i=1}^{I} a_{ij}}, \tag{12}$$

$$a_{ij} \leftarrow a_{ij} \frac{\sum_{t=1}^{T} x_{jt} \left(p_{it}/q_{it}\right)}{\sum_{t=1}^{T} x_{jt}}. \tag{13}$$

We shall derive the above algorithm in a much more general and flexible form in Section 3.

To accelerate the convergence of the EMML, we can apply the following heuristic extension of the EMML algorithm [8,51]:

$$x_{jt} \leftarrow x_{jt} \left(\frac{\sum_{i=1}^{I} a_{ij} \left(p_{it}/q_{it}\right)}{\sum_{i=1}^{I} a_{ij}} \right)^{w}, \tag{14}$$

$$a_{ij} \leftarrow a_{ij} \left(\frac{\sum_{t=1}^{T} x_{jt} \left(p_{it}/q_{it}\right)}{\sum_{t=1}^{T} x_{jt}} \right)^{w}, \tag{15}$$

where the positive relaxation parameter w helps to improve the convergence (see Section 3.3).

In an enhanced form to reduce the computational cost, the denominators in (14) and (15) can be ignored due to normalizing $\boldsymbol{a}_j$ to the unit length of ℓ_1-norm:

$$\mathbf{X} \leftarrow \mathbf{X} \circledast \left(\mathbf{A}^T \left(\mathbf{P} \oslash \mathbf{Q}\right)\right)^{.[w]}, \tag{16}$$

$$\mathbf{A} \leftarrow \mathbf{A} \circledast \left(\left(\mathbf{P} \oslash \mathbf{Q}\right) \mathbf{X}^T\right)^{.[w]}, \tag{17}$$

$$a_{ij} \leftarrow a_{ij}/\|\boldsymbol{a}_j\|_1. \tag{18}$$

One of the objectives of this paper is develop generalized multiplicative NMF algorithms which are robust with respect to noise and/or outliers and integrate (combine) both the ISRA and the EMML algorithms into the more general a flexible one.

2. The Alpha-Beta Divergences

For positive measures $\mathbf{P}$ and $\mathbf{Q}$ consider the following new dissimilarity measure, which we shall refer to as the AB-divergence:

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = -\frac{1}{\alpha\beta} \sum_{it} \left(p_{it}^{\alpha} q_{it}^{\beta} - \frac{\alpha}{\alpha+\beta} p_{it}^{\alpha+\beta} - \frac{\beta}{\alpha+\beta} q_{it}^{\alpha+\beta} \right) \tag{19}$$
$$\text{for } \alpha, \beta, \alpha+\beta \neq 0,$$

or equivalently

$$D_{AB}^{(\alpha,\lambda-\alpha)}(\mathbf{P}\|\mathbf{Q}) = \frac{1}{(\alpha-\lambda)\alpha}\sum_{it}\left(p_{it}^{\alpha}q_{it}^{\lambda-\alpha} - \frac{\alpha}{\lambda}p_{it}^{\lambda} - \frac{\lambda-\alpha}{\lambda}q_{it}^{\lambda}\right), \quad \text{for } \alpha \neq 0,\ \alpha\neq\lambda,\ \lambda = \alpha+\beta \neq 0, \tag{20}$$

Note that, Equation (19) is a divergence since the following relationship holds:

$$\frac{1}{\alpha\beta}p_{it}^{\alpha}q_{it}^{\beta} \leq \frac{1}{\beta(\alpha+\beta)}p_{it}^{\alpha+\beta} + \frac{1}{\alpha(\alpha+\beta)}q_{it}^{\alpha+\beta}, \quad \text{for } \alpha,\beta,\alpha+\beta\neq 0, \tag{21}$$

with equality holding for $p_{it} = q_{it}$. In Appendix A, we show that Equation (21) is a summarization of three different inequalities of Young's type, each one holding true for a different combination of the signs of the constants: $\alpha\beta$, $\alpha(\alpha+\beta)$ and $\beta(\alpha+\beta)$.

In order to avoid indeterminacy or singularity for certain values of parameters, the AB-divergence can be extended by continuity (by applying l'Hôpital formula) to cover all the values of $\alpha,\beta\in\mathbb{R}$, thus the AB-divergence can be expressed or defined in a more explicit form:

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = \sum_{it} d_{AB}^{(\alpha,\beta)}(p_{it},q_{it}), \tag{22}$$

where

$$d_{AB}^{(\alpha,\beta)}(p_{it},q_{it}) = \begin{cases} -\frac{1}{\alpha\beta}\left(p_{it}^{\alpha}q_{it}^{\beta} - \frac{\alpha}{(\alpha+\beta)}p_{it}^{\alpha+\beta} - \frac{\beta}{(\alpha+\beta)}q_{it}^{\alpha+\beta}\right) & \text{for } \alpha,\beta,\alpha+\beta\neq 0 \\ \frac{1}{\alpha^2}\left(p_{it}^{\alpha}\ln\frac{p_{it}^{\alpha}}{q_{it}^{\alpha}} - p_{it}^{\alpha} + q_{it}^{\alpha}\right) & \text{for } \alpha\neq 0, \beta = 0 \\ \frac{1}{\alpha^2}\left(\ln\frac{q_{it}^{\alpha}}{p_{it}^{\alpha}} + \left(\frac{q_{it}^{\alpha}}{p_{it}^{\alpha}}\right)^{-1} - 1\right) & \text{for } \alpha = -\beta\neq 0 \\ \frac{1}{\beta^2}\left(q_{it}^{\beta}\ln\frac{q_{it}^{\beta}}{p_{it}^{\beta}} - q_{it}^{\beta} + p_{it}^{\beta}\right) & \text{for } \alpha = 0, \beta\neq 0 \\ \frac{1}{2}\left(\ln p_{it} - \ln q_{it}\right)^2 & \text{for } \alpha,\beta = 0. \end{cases} \tag{23}$$

2.1. Special Cases of the AB-Divergence

We shall now illustrate that a suitable choice of the (α,β) parameters simplifies the AB-divergence into some existing divergences, including the well-known Alpha- and Beta-divergences [4,7,19,35].

When $\alpha+\beta = 1$ the AB-divergence reduces to the Alpha-divergence [4,33–35,52]

$$D_{AB}^{(\alpha,1-\alpha)}(\mathbf{P}\|\mathbf{Q}) = D_{A}^{(\alpha)}(\mathbf{P}\|\mathbf{Q}) \tag{24}$$

$$\doteq \begin{cases} \frac{1}{\alpha(\alpha-1)}\sum_{it}\left(p_{it}^{\alpha}q_{it}^{1-\alpha} - \alpha p_{it} + (\alpha-1)q_{it}\right) & \text{for } \alpha\neq 0, \alpha\neq 1, \\ \sum_{it}\left(p_{it}\ln\frac{p_{it}}{q_{it}} - p_{it} + q_{it}\right) & \text{for } \alpha = 1, \\ \sum_{it}\left(q_{it}\ln\frac{q_{it}}{p_{it}} - q_{it} + p_{it}\right) & \text{for } \alpha = 0. \end{cases} \tag{25}$$

On the other hand, when $\alpha = 1$, it reduces to the Beta-divergence [8,9,19,32,53]

$$D_{AB}^{(1,\beta)}(\mathbf{P}\|\mathbf{Q}) = D_B^{(\beta)}(\mathbf{P}\|\mathbf{Q}) \tag{26}$$

$$\doteq \begin{cases} -\dfrac{1}{\beta}\displaystyle\sum_{it}\left(p_{it}q_{it}^{\beta} - \dfrac{1}{(1+\beta)}p_{it}^{1+\beta} - \dfrac{\beta}{(1+\beta)}q_{it}^{1+\beta}\right) & \text{for } \beta, 1+\beta \neq 0, \\ \dfrac{1}{2}\displaystyle\sum_{it}(p_{it}-q_{it})^2 & \text{for } \beta = 1, \\ \displaystyle\sum_{it}\left(p_{it}\ln\dfrac{p_{it}}{q_{it}} - p_{it} + q_{it}\right) & \text{for } \beta = 0, \\ \displaystyle\sum_{it}\left(\ln\dfrac{q_{it}}{p_{it}} + \left(\dfrac{q_{it}}{p_{it}}\right)^{-1} - 1\right) & \text{for } \beta = -1. \end{cases} \tag{27}$$

The AB-divergence gives to the standard Kullback-Leibler (KL) divergence ($D_{KL}(\cdot,\cdot)$) for $\alpha = 1$ and $\beta = 0$

$$D_{AB}^{(1,0)}(\mathbf{P}\|\,\mathbf{Q}) = D_{KL}(\mathbf{P}\|\,\mathbf{Q}) = \sum_{it}\left(p_{it}\ln\frac{p_{it}}{q_{it}} - p_{it} + q_{it}\right), \tag{28}$$

and reduces to the standard Itakura-Saito divergence ($D_{IS}(\cdot,\cdot)$) for $\alpha = 1$ and $\beta = -1$ [8,53,54]

$$D_{AB}^{(1,-1)}(\mathbf{P}\|\,\mathbf{Q}) = D_{IS}(\mathbf{P}\|\,\mathbf{Q}) = \sum_{it}\left(\ln\frac{q_{it}}{p_{it}} + \frac{p_{it}}{q_{it}} - 1\right). \tag{29}$$

Using the $1-\alpha$ deformed logarithm defined as

$$\ln_{1-\alpha}(z) = \begin{cases} \dfrac{z^{\alpha}-1}{\alpha}, & \alpha \neq 0, \\ \ln z, & \alpha = 0, \end{cases} \tag{30}$$

where $z > 0$, observe that the AB-divergence is symmetric with respect to both arguments for $\alpha = \beta \neq 0$ and takes the form of a metric distance

$$D_{AB}^{(\alpha,\alpha)}(\mathbf{P}\|\mathbf{Q}) = D_E(\ln_{1-\alpha}(\mathbf{P})\|\ln_{1-\alpha}(\mathbf{Q})) = \frac{1}{2}\sum_{it}\left(\ln_{1-\alpha}(p_{it}) - \ln_{1-\alpha}(q_{it})\right)^2, \tag{31}$$

in the transform domain $\phi(x) = \ln_{1-\alpha}(x)$. As particular cases, this includes the scaled squared Euclidean distance (for $\alpha = 1$) and the Hellinger distance (for $\alpha = 0.5$).

For both $\alpha \to 0$ and $\beta \to 0$, the AB-divergence converges to the Log-Euclidean distance defined as

$$D_{AB}^{(0,0)}(\mathbf{P}\|\mathbf{Q}) = \lim_{\alpha\to 0} D_{AB}^{(\alpha,\alpha)}(\mathbf{P}\|\mathbf{Q}) = \frac{1}{2}\sum_{it}(\ln p_{it} - \ln q_{it})^2. \tag{32}$$

2.2. *Properties of AB-Divergence: Duality, Inversion and Scaling*

Let us denote by $\mathbf{P}^{\cdot[r]}$ the one to one transformation that raises each positive element of the matrix $\mathbf{P}$ to the power r, *i.e.*, each entry is raised to power r, that is p_{it}^r. According to the definition of the AB-divergence, we can easily check that it satisfies the following duality property (see Figure 1)

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = D_{AB}^{(\beta,\alpha)}(\mathbf{Q}\|\mathbf{P}), \tag{33}$$

and inversion property

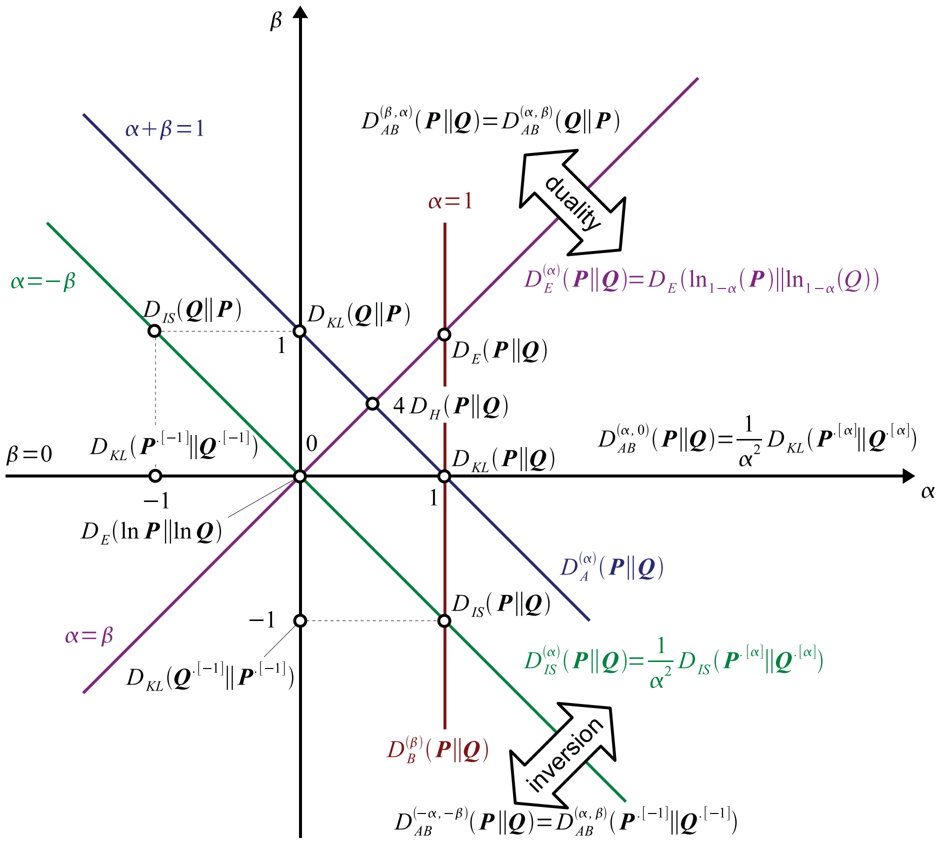
$$D_{AB}^{(-\alpha,-\beta)}(\mathbf{P}\|\mathbf{Q}) \;=\; D_{AB}^{(\alpha,\beta)}(\mathbf{P}^{\cdot\,[-1]})\|\mathbf{Q}^{\cdot\,[-1]}) \,. \tag{34}$$

The above properties can be considered as a particular case of the scaling property of parameters α and β by a common factor $\omega \in \mathbb{R} \setminus \{0\}$. The divergence whose parameters has been rescaled is proportional to original divergence with both arguments raised to the common factor, *i.e.*,

$$D_{AB}^{(\omega\alpha,\omega\beta)}(\mathbf{P}\|\mathbf{Q}) \;=\; \frac{1}{\omega^2} D_{AB}^{(\alpha,\beta)}(\mathbf{P}^{\cdot\,[\omega]} \,\|\, \mathbf{Q}^{\cdot\,[\omega]}) \,. \tag{35}$$

The scaling of α and β by a common factor $\omega < 1$ can be seen as a 'zoom-in' on the arguments $\mathbf{P}$ and $\mathbf{Q}$. This zoom gives more relevance to the smaller values over the large values. On the other hand, for $\omega > 1$ yields a 'zoom-out' effect were the smaller values decrease their relevance at the expense of large values whose relevance is increased (see Figure 1).

Figure 1. Graphical illustration of duality and inversion properties of the AB-divergence. On the alpha-beta plane are indicated as special important cases particular divergences by points and lines, especially Kullback-Leibler divergence D_{KL}, Hellinger Distance D_H, Euclidean distance D_E, Itakura-Saito distance D_{IS}, Alpha-divergence $D_A^{(\alpha)}$, and Beta-divergence $D_B^{(\beta)}$.


β
$\alpha+\beta=1$
$D_{AB}^{(\beta,\alpha)}(\boldsymbol{P}\|\boldsymbol{Q})=D_{AB}^{(\alpha,\beta)}(\boldsymbol{Q}\|\boldsymbol{P})$
duality
$\alpha=1$
$\alpha=-\beta$
$D_E^{(\alpha)}(\boldsymbol{P}\|\boldsymbol{Q})=D_E(\ln_{1-\alpha}(\boldsymbol{P})\|\ln_{1-\alpha}(\boldsymbol{Q}))$
$D_{IS}(\boldsymbol{Q}\|\boldsymbol{P})$
$D_{KL}(\boldsymbol{Q}\|\boldsymbol{P})$
1
$D_E(\boldsymbol{P}\|\boldsymbol{Q})$
$4\,D_H(\boldsymbol{P}\|\boldsymbol{Q})$
$\beta=0$
$D_{KL}(\boldsymbol{P}^{\cdot[-1]}\|\boldsymbol{Q}^{\cdot[-1]})$
0
$D_{AB}^{(\alpha,0)}(\boldsymbol{P}\|\boldsymbol{Q})=\frac{1}{\alpha^2}D_{KL}(\boldsymbol{P}^{\cdot[\alpha]}\|\boldsymbol{Q}^{\cdot[\alpha]})$
$D_{KL}(\boldsymbol{P}\|\boldsymbol{Q})$
α
-1
1
$D_E(\ln\boldsymbol{P}\|\ln\boldsymbol{Q})$
$D_A^{(\alpha)}(\boldsymbol{P}\|\boldsymbol{Q})$
$D_{IS}(\boldsymbol{P}\|\boldsymbol{Q})$
-1
$\alpha=\beta$
$D_{IS}^{(\alpha)}(\boldsymbol{P}\|\boldsymbol{Q})=\frac{1}{\alpha^2}D_{IS}(\boldsymbol{P}^{\cdot[\alpha]}\|\boldsymbol{Q}^{\cdot[\alpha]})$
$D_{KL}(\boldsymbol{Q}^{\cdot[-1]}\|\boldsymbol{P}^{\cdot[-1]})$
$D_B^{(\beta)}(\boldsymbol{P}\|\boldsymbol{Q})$
inversion
$D_{AB}^{(-\alpha,-\beta)}(\boldsymbol{P}\|\boldsymbol{Q})=D_{AB}^{(\alpha,\beta)}(\boldsymbol{P}^{\cdot[-1]}\|\boldsymbol{Q}^{\cdot[-1]})$

By scaling arguments of the AB-divergence by a positive scaling factor $c > 0$, it yields the following relation

$$D_{AB}^{(\alpha,\beta)}(c\,\mathbf{P}\|c\,\mathbf{Q}) \;=\; c^{\alpha+\beta}\, D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\,\mathbf{Q}) \,. \tag{36}$$

These basic properties imply that whenever $\alpha \neq 0$, we can rewrite the AB-divergence in terms of a $\frac{\beta}{\alpha}$-order Beta-divergence combined with an α-zoom of its arguments as

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\,\mathbf{Q}) = \frac{1}{\alpha^2} D_{AB}^{\left(1,\frac{\beta}{\alpha}\right)}(\mathbf{P}^{.\,[\alpha]}\|\,\mathbf{Q}^{.\,[\alpha]}) \tag{37}$$

$$= \frac{1}{\alpha^2} D_{B}^{\left(\frac{\beta}{\alpha}\right)}(\mathbf{P}^{.\,[\alpha]}\|\,\mathbf{Q}^{.\,[\alpha]})\,. \tag{38}$$

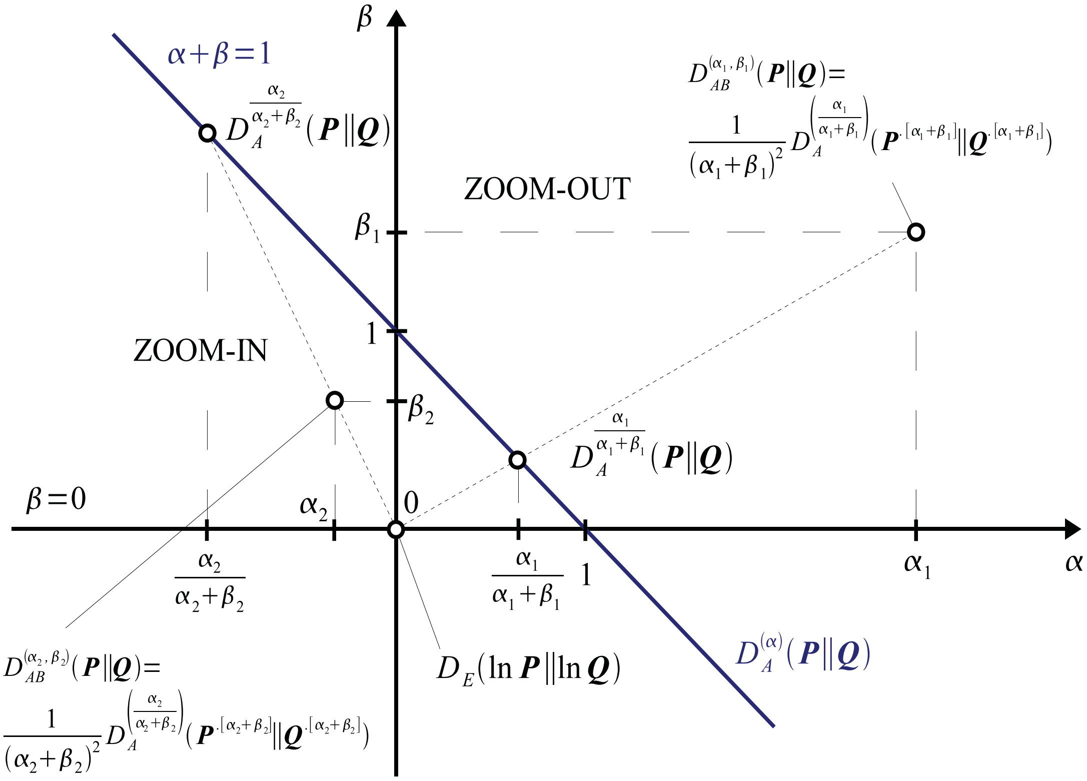
Similarly, for $\alpha + \beta \neq 0$, the AB-divergence can also be expressed in terms of $\frac{\alpha}{\alpha+\beta}$-order Alpha-divergence with an $(\alpha + \beta)$-zoom of the arguments (see Figure 2) as

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\,\mathbf{Q}) = \frac{1}{(\alpha+\beta)^2} D_{AB}^{\left(\frac{\alpha}{\alpha+\beta},\frac{\beta}{\alpha+\beta}\right)}(\mathbf{P}^{.\,[\alpha+\beta]}\|\,\mathbf{Q}^{.\,[\alpha+\beta]}) \tag{39}$$

$$= \frac{1}{(\alpha+\beta)^2} D_{A}^{\left(\frac{\alpha}{\alpha+\beta}\right)}\left(\mathbf{P}^{.\,[\alpha+\beta]}\|\,\mathbf{Q}^{.\,[\alpha+\beta]}\right)\,. \tag{40}$$

illustrating that the AB-divergence equals to an $\frac{\alpha}{\alpha+\beta}$-order Alpha-divergence with an $(\alpha+\beta)$-zoom in of the arguments. On the other hand $D_{AB}^{(\alpha_1,\beta_1)}$ can be seen as an Alpha-divergence of order $\alpha_1/(\alpha_1 + \beta_1)$ of a zoom-out of the arguments (with $\alpha_1+\beta_1 > 1$). Note that, $D_{AB}^{(\alpha_2,\beta_2)}$ can be seen as the Alpha-divergence of order $\alpha_2/(\alpha_2 + \beta_2)$ with a zoom-in of its arguments with $\alpha_2 + \beta_2 < 1$ (see Figure 2).

Figure 2. Illustrations how the AB-divergence for $\alpha + \beta \neq 0$ can be expressed via scaled Alpha-divergences.


β
$\alpha+\beta=1$
$D_A^{\frac{\alpha_2}{\alpha_2+\beta_2}}(\boldsymbol{P}\|\boldsymbol{Q})$
$D_{AB}^{(\alpha_1,\beta_1)}(\boldsymbol{P}\|\boldsymbol{Q})=$
$\frac{1}{(\alpha_1+\beta_1)^2}D_A^{\left(\frac{\alpha_1}{\alpha_1+\beta_1}\right)}(\boldsymbol{P}^{.[\alpha_1+\beta_1]}\|\boldsymbol{Q}^{.[\alpha_1+\beta_1]})$
ZOOM-OUT
β_1
1
ZOOM-IN
β_2
$D_A^{\frac{\alpha_1}{\alpha_1+\beta_1}}(\boldsymbol{P}\|\boldsymbol{Q})$
$\beta=0$
α_2
0
$\frac{\alpha_2}{\alpha_2+\beta_2}$
$\frac{\alpha_1}{\alpha_1+\beta_1}$
1
α_1
α
$D_{AB}^{(\alpha_2,\beta_2)}(\boldsymbol{P}\|\boldsymbol{Q})=$
$\frac{1}{(\alpha_2+\beta_2)^2}D_A^{\left(\frac{\alpha_2}{\alpha_2+\beta_2}\right)}(\boldsymbol{P}^{.[\alpha_2+\beta_2]}\|\boldsymbol{Q}^{.[\alpha_2+\beta_2]})$
$D_E(\ln\boldsymbol{P}\|\ln\boldsymbol{Q})$
$D_A^{(\alpha)}(\boldsymbol{P}\|\boldsymbol{Q})$

When $\alpha \neq 0$ and $\beta = 0$, the AB-divergence can be expressed in terms of the Kullback-Leibler divergence with an α-zoom of its arguments

$$\begin{aligned} D_{AB}^{(\alpha,0)}(\mathbf{P}\|\,\mathbf{Q}) &= \frac{1}{\alpha^2} D_{AB}^{(1,0)}\left(\mathbf{P}^{\cdot\,[\alpha]}\|\,\mathbf{Q}^{\cdot\,[\alpha]}\right) && (41)\\ &= \frac{1}{\alpha^2} D_{KL}\left(\mathbf{P}^{\cdot\,[\alpha]}\|\,\mathbf{Q}^{\cdot\,[\alpha]}\right) && (42)\\ &= \frac{1}{\alpha^2}\sum_{it}\left(p_{it}^{\alpha}\ln\frac{p_{it}^{\alpha}}{q_{it}^{\alpha}} - p_{it}^{\alpha} + q_{it}^{\alpha}\right). && (43)\end{aligned}$$

When $\alpha + \beta = 0$ with $\alpha \neq 0$ and $\beta \neq 0$, the AB-divergence can also be expressed in terms of a generalized Itakura-Saito distance with an α-zoom of the arguments

$$\begin{aligned} D_{AB}^{(\alpha,-\alpha)}(\mathbf{P}\|\,\mathbf{Q}) &= \frac{1}{\alpha^2} D_{AB}^{(1,-1)}(\mathbf{P}^{\cdot\,[\alpha]}\|\,\mathbf{Q}^{\cdot\,[\alpha]}) && (44)\\ &= \frac{1}{\alpha^2} D_{IS}\left(\mathbf{P}^{\cdot\,[\alpha]}\|\,\mathbf{Q}^{\cdot\,[\alpha]}\right) && (45)\\ &= \frac{1}{\alpha^2}\sum_{it}\left(\ln\frac{q_{it}^{\alpha}}{p_{it}^{\alpha}} + \frac{p_{it}^{\alpha}}{q_{it}^{\alpha}} - 1\right). && (46)\end{aligned}$$

Remark: The generalized Itakura-Saito distance is scale-invariant, *i.e*,

$$D_{AB}^{(\alpha,-\alpha)}(\mathbf{P}\|\,\mathbf{Q}) = D_{AB}^{(\alpha,-\alpha)}(c\mathbf{P}\|\,c\mathbf{Q}), \tag{47}$$

with any $c > 0$.

Note that, from the AB-divergence, a more general scale-invariant divergence can be obtained by the monotonic nonlinear transformations:

$$\sum_{it} p_{it}^{\alpha} q_{it}^{\beta} \to \ln\left(\sum_{it} p_{it}^{\alpha} q_{it}^{\beta}\right) \qquad \forall\, \alpha, \beta \tag{48}$$

leading to the following scale-invariant divergence given by

$$\begin{aligned} D_{AC}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) &= \frac{1}{\beta(\alpha+\beta)}\ln\left(\sum_{it} p_{it}^{\alpha+\beta}\right) + \frac{1}{\alpha(\alpha+\beta)}\ln\left(\sum_{it} q_{it}^{\alpha+\beta}\right) - \frac{1}{\alpha\beta}\ln\left(\sum_{it} p_{it}^{\alpha} q_{it}^{\beta}\right)\\ &= \frac{1}{\alpha\beta}\ln\frac{\left(\sum_{it} p_{it}^{\alpha+\beta}\right)^{\frac{\alpha}{\alpha+\beta}}\left(\sum_{it} q_{it}^{\alpha+\beta}\right)^{\frac{\beta}{\alpha+\beta}}}{\sum_{it} p_{it}^{\alpha} q_{it}^{\beta}} \quad \text{for} \quad \alpha \neq 0,\ \beta \neq 0,\ \alpha+\beta \neq 0. \quad (49)\end{aligned}$$

which generalizes a family of Gamma-divergences [35].

The divergence (49) is scale-invariant in a more general context

$$D_{AC}^{(\alpha,\beta)}(\mathbf{P}||\mathbf{Q}) = D_{AC}^{(\alpha,\beta)}(c_1\,\mathbf{P}||c_2\,\mathbf{Q}) \tag{50}$$

for any positive scale factors c_1 and c_2.

2.3. Why is AB-Divergence Potentially Robust?

To illustrate the role of the hyperparameters α and β on the robustness of the AB-divergence with respect to errors and noises, we shall compare the behavior of the AB-divergence with the standard Kullback-Leibler divergence. We shall assume, without loss of generality, that the proposed factorization model $\mathbf{Q}$ (for given noisy $\mathbf{P}$) is a function of the vector of parameters $\boldsymbol{\theta}$ and that each of its elements $q_{it}(\boldsymbol{\theta}) > 0$ is non-negative for a certain range of parameters $\boldsymbol{\Theta}$.

The estimator $\hat{\boldsymbol{\theta}}$ obtained for the Kullback-Leibler divergence between two discrete positive measures $\mathbf{P}$ and $\mathbf{Q}$, is a solution of

$$\frac{\partial D_{KL}(\mathbf{P}\|\mathbf{Q})}{\partial \boldsymbol{\theta}} = -\sum_{it} \frac{\partial q_{it}}{\partial \boldsymbol{\theta}} \ln_0\left(\frac{p_{it}}{q_{it}}\right) = \mathbf{0}\,, \tag{51}$$

while, for the Beta-divergence, the estimator solves

$$\frac{\partial D_B^{(\beta)}(\mathbf{P}\|\mathbf{Q})}{\partial \boldsymbol{\theta}} = -\sum_{it} \frac{\partial q_{it}}{\partial \boldsymbol{\theta}}\, q_{it}^{\beta} \ln_0\left(\frac{p_{it}}{q_{it}}\right) = \mathbf{0}\,. \tag{52}$$

The main difference between both equations is in the weighting factors q_{it}^{β} for the Beta-divergence which are controlled by the parameter β. In the context of probability distributions, these weighting factors may control the influence of likelihood ratios p_{it}/q_{it}. It has been shown [55] and [56] that the parameter β determines a tradeoff between robustness to outliers (for $\beta > 0$) and efficiency (for β near 0). In the special case of $\beta = 1$ the Euclidean distance is obtained, which is known to be more robust and less efficient than the Kullback-Leibler divergence $\beta = 0$.

On the other hand, for the Alpha-divergence, the estimating equation takes a different form

$$\frac{\partial D_A^{(\alpha)}(\mathbf{P}\|\mathbf{Q})}{\partial \boldsymbol{\theta}} = -\sum_{it} \frac{\partial q_{it}}{\partial \boldsymbol{\theta}} \ln_{1-\alpha}\left(\frac{p_{it}}{q_{it}}\right) = \mathbf{0}\,. \tag{53}$$

In this case, the influence of the values of individual ratios p_{it}/q_{it} are controlled not by weighting factors but by the deformed logarithm of order $1-\alpha$. This feature can be interpreted as a zoom or over-weight of the interesting details of the likelihood ratio. For $\alpha > 1$ (a zoom-out), we emphasize the relative importance of larger values of the ratio p_{it}/q_{it}, whereas for $\alpha < 1$ (a zoom-in), we put more emphasis on smaller values of p_{it}/q_{it} (see Figure 3). The major consequence is the inclusive ($\alpha \to \infty$) and exclusive ($\alpha \leftarrow -\infty$) behavior of the Alpha-divergence discussed in [52].

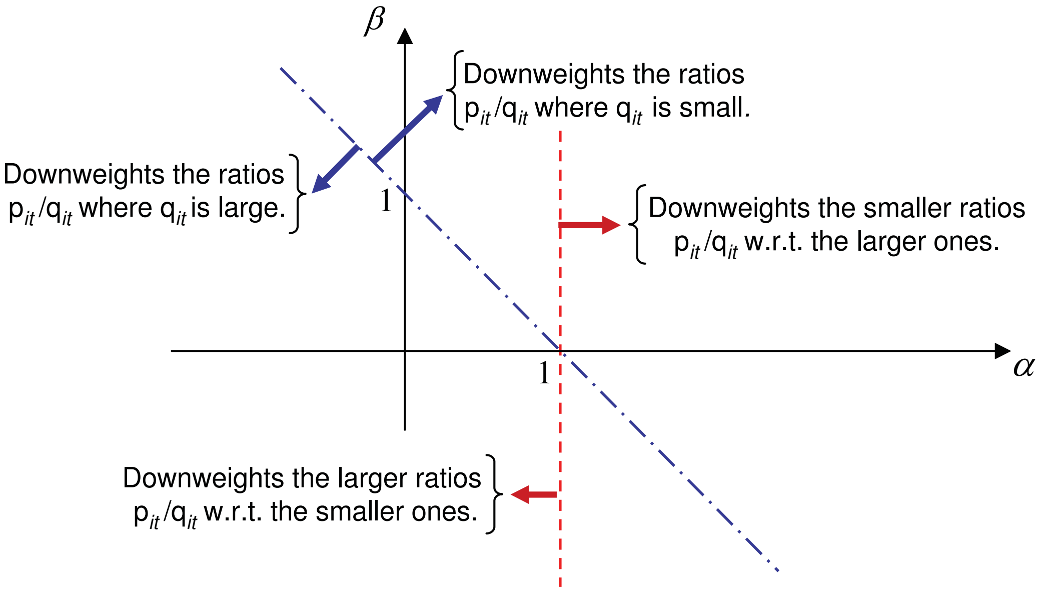
The estimating equation for the Alpha-Beta divergence combines both effects:

$$\frac{\partial D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})}{\partial \boldsymbol{\theta}} = -\sum_{it} \frac{\partial q_{it}}{\partial \boldsymbol{\theta}} \underbrace{q_{it}^{\alpha+\beta-1}}_{weights} \underbrace{\ln_{1-\alpha}(p_{it}/q_{it})}_{\alpha-zoom} = \mathbf{0}\,, \tag{54}$$

and therefore is much more flexible and powerful regarding insensitivity to error and noise.

As illustrated in Figure 3, depending on value of α, we can zoom-in or zoom-out the interesting sets of the ratios p_{it}/q_{it} and simultaneously weight these ratios by scaling factors $q_{it}^{\lambda-1}$ controlled by the parameter $\lambda = \alpha + \beta$. Therefore, the parameter α can be used to control the influence of large or small ratios in the estimator, while the parameter β provides some control on the weighting of the ratios depending on the demand to better fit to larger or smaller values of the model.

Figure 3. Graphical illustration how the set parameters (α, β) can control influence of individual ratios p_{it}/q_{it}. The dash-doted line ($\alpha + \beta = 1$) shows the region where the multiplicative weighting factor $q_{it}^{\alpha+\beta-1}$ in the estimating equations is constant and equal to unity. The dashed line ($\alpha = 1$) shows the region where the order of the deformed logarithm of p_{it}/q_{it} is constant and equal to that of the standard Kullback-Leibler divergence.


β
Downweights the ratios p_{it}/q_{it} where q_{it} is small.
Downweights the ratios p_{it}/q_{it} where q_{it} is large.
1
Downweights the smaller ratios p_{it}/q_{it} w.r.t. the larger ones.
1
α
Downweights the larger ratios p_{it}/q_{it} w.r.t. the smaller ones.

3. Generalized Multiplicative Algorithms for NMF

3.1. Derivation of Multiplicative NMF Algorithms Based on the AB-Divergence

We shall now develop generalized and flexible NMF algorithms (see Equation (1)) by employing the AB-divergence with $\mathbf{Y} = \mathbf{P} = [p_{it}] \in \mathbb{R}_+^{I\times T}$, $\mathbf{Q} = [q_{it}] = \mathbf{AX} \in \mathbb{R}_+^{I\times T}$, where $q_{it} = \widehat{p}_{it} = [\mathbf{AX}]_{it} = \sum_j a_{ij}x_{jt}$. In this case, the gradient of the AB-divergence (20) can be expressed in a compact form (for any $\alpha, \beta \in \mathbb{R}$) in terms of an $1 - \alpha$ deformed logarithm (see (30))

$$\frac{\partial D_{AB}^{(\alpha,\beta)}}{\partial x_{jt}} = -\sum_{i=1}^{I} q_{it}^{\lambda-1}\, a_{ij} \ln_{1-\alpha}\left(\frac{p_{it}}{q_{it}}\right), \qquad (55)$$

$$\frac{\partial D_{AB}^{(\alpha,\beta)}}{\partial a_{ij}} = -\sum_{t=1}^{T} q_{it}^{\lambda-1}\, x_{jt} \ln_{1-\alpha}\left(\frac{p_{it}}{q_{it}}\right). \qquad (56)$$

The new multiplicative learning algorithm is gradient descent based, in the natural parameter space $\phi(x)$ of the proposed divergence:

$$x_{jt} \leftarrow \phi^{-1}\left(\phi(x_{jt}) - \eta_{jt}\frac{\partial D_{AB}^{(\alpha,\beta)}}{\partial \phi(x_{jt})}\right), \qquad (57)$$

$$a_{ij} \leftarrow \phi^{-1}\left(\phi(a_{ij}) - \eta_{ij}\frac{\partial D_{AB}^{(\alpha,\beta)}}{\partial \phi(a_{ij})}\right). \qquad (58)$$

These updates can be considered as a generalization of the exponentiated gradient (EG) [57].

In general, such a nonlinear scaling (or transformation) provides a stable solution and the obtained gradients are much better behaved in the ϕ space. We use the $1-\alpha$ deformed logarithm transformation $\phi(z) = \ln_{1-\alpha}(z)$, whose inverse transformation is a $1-\alpha$ deformed exponential

$$\phi^{-1}(z) = \exp_{1-\alpha}(z) = \begin{cases} \exp(z) & \text{for} \quad \alpha = 0, \\ (1+\alpha z)^{\frac{1}{\alpha}} & \text{for} \quad \alpha \neq 0 \ \text{ and } 1+\alpha z \geq 0, \\ 0 & \text{for} \quad \alpha \neq 0 \ \text{ and } 1+\alpha z < 0. \end{cases} \tag{59}$$

For positive measures $z > 0$, the direct transformation $\phi(z)$ and the composition $\phi^{-1}(\phi(z))$ are bijective functions which define a one to one correspondence, so we have $\phi^{-1}(\phi(z)) = z$.

By choosing suitable learning rates

$$\eta_{jt} = \frac{x_{jt}^{2\alpha-1}}{\sum\limits_{i=1}^{I} a_{ij} q_{it}^{\lambda-1}}, \qquad \eta_{ij} = \frac{a_{ij}^{2\alpha-1}}{\sum\limits_{t=1}^{T} x_{jt} q_{it}^{\lambda-1}}. \tag{60}$$

a new multiplicative NMF algorithm (refereed to as the AB-multiplicative NMF algorithm) is obtained as:

$$\begin{aligned} x_{jt} &\leftarrow x_{jt} \; \exp_{1-\alpha}\left(\sum_{i=1}^{I} \frac{a_{ij}\, q_{it}^{\lambda-1}}{\sum_{i=1}^{I} a_{ij}\, q_{it}^{\lambda-1}} \; \ln_{1-\alpha}\left(\frac{p_{it}}{q_{it}}\right) \right), \\ a_{ij} &\leftarrow a_{ij} \; \exp_{1-\alpha}\left(\sum_{t=1}^{T} \frac{x_{jt}\, q_{it}^{\lambda-1}}{\sum_{t=1}^{t} x_{jt}\, q_{it}^{\lambda-1}} \; \ln_{1-\alpha}\left(\frac{p_{it}}{q_{it}}\right) \right). \end{aligned} \tag{61}$$

In these equations, the deformed logarithm of order $1-\alpha$ of the quotients p_{it}/q_{it} plays a key role to control relative error terms whose weighted mean provides the multiplicative corrections. This deformation in the relative error is controlled by the parameter α. The parameter $\alpha > 1$ gives more relevance to the large values of the quotient, while the case of $\alpha < 1$ puts more emphasis on smaller values of the quotient. On the other hand, the parameter $\lambda - 1$ (where $\lambda = \alpha + \beta$) controls the influence of the values of the approximation (q_{it}) on the weighing of the deformed error terms. For $\lambda = 1$, this influence disappears.

It is interesting to note that the multiplicative term of the main updates

$$M_\alpha(\mathbf{z}, \mathbf{w}, S) \;=\; \exp_{1-\alpha}\left(\frac{1}{\sum\limits_{i \in S} w_i} \sum_{i \in S} w_i \; \ln_{1-\alpha}(z_i) \right), \tag{62}$$

can be interpreted as a weighted generalized mean across the elements with indices in the set S.

Depending on the value of α, we obtain as particular cases: the minimum of the vector $\mathbf{z}$ (for $\alpha \to -\infty$), its weighted harmonic mean ($\alpha = -1$), the weighted geometric mean ($\alpha = 0$), the arithmetic mean ($\alpha = 1$), the weighted quadratic mean ($\alpha = 2$) and the maximum of the vector ($\alpha \to \infty$), *i.e.*,

$$M_\alpha(\mathbf{z}, \mathbf{w}, \{1, \ldots, n\}) = \begin{cases} \min\{z_1, \ldots, z_n\}, & \alpha \to -\infty, \\ \left(\sum_{i=1}^{n} w_i\right)\left(\sum_{i=1}^{n} \frac{w_i}{z_i}\right)^{-1}, & \alpha = -1, \\ \prod_{i=1}^{n} z^{\frac{w_i}{\sum_{i=1}^{n} w_i}}, & \alpha = 0, \\ \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i\, z_i, & \alpha = 1, \\ \left(\frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i\, z_i^2\right)^{1/2}, & \alpha = 2, \\ \max\{z_1, \ldots, z_n\}, & \alpha \to \infty. \end{cases} \tag{63}$$

The generalized weighted means are monotonically increasing functions of α, *i.e.*, if $\alpha_1 < \alpha_2$, then

$$M_{\alpha_1}(\mathbf{z}, \mathbf{w}, S) < M_{\alpha_2}(\mathbf{z}, \mathbf{w}, S)\,. \tag{64}$$

Thus, by increasing the values of α, we puts more emphasis on large relative errors in the update formulas (61).

In the special case of $\alpha \neq 0$, the above update rules can be simplified as:

$$x_{jt} \leftarrow x_{jt} \left(\frac{\sum_{i=1}^{I} a_{ij}\, p_{it}^{\alpha}\, q_{it}^{\beta-1}}{\sum_{i=1}^{I} a_{ij}\, q_{it}^{\alpha+\beta-1}} \right)^{1/\alpha}, \tag{65}$$

$$a_{ij} \leftarrow a_{ij} \left(\frac{\sum_{t=1}^{T} x_{jt}\, p_{it}^{\alpha}\, q_{it}^{\beta-1}}{\sum_{t=1}^{T} x_{jt}\, q_{it}^{\alpha+\beta-1}} \right)^{1/\alpha}, \tag{66}$$

where $q_{it} = [\mathbf{A}\mathbf{X}]_{it}$ and at every iteration the columns of $\mathbf{A}$ are normalized to the unit length.

The above multiplicative update rules can be written in a compact matrix forms as

$$\mathbf{X} \leftarrow \mathbf{X} \circledast \left(\left(\mathbf{A}^T (\mathbf{P}^{\cdot[\alpha]} \circledast \mathbf{Q}^{\cdot[\beta-1]}) \right) \oslash \left(\mathbf{A}^T \mathbf{Q}^{\cdot[\alpha+\beta-1]} \right) \right)^{\cdot[1/\alpha]}, \tag{67}$$

$$\mathbf{A} \leftarrow \mathbf{A} \circledast \left(\left((\mathbf{P}^{\cdot[\alpha]} \circledast \mathbf{Q}^{\cdot[\beta-1]})\, \mathbf{X}^T \right) \oslash \left(\mathbf{Q}^{\cdot[\alpha+\beta-1]}\, \mathbf{X}^T \right) \right)^{\cdot[1/\alpha]} \tag{68}$$

or even more compactly:

$$\mathbf{X} \leftarrow \mathbf{X} \circledast \left(\left(\mathbf{A}^T \mathbf{Z} \right) \oslash \left(\mathbf{A}^T \mathbf{Q}^{\cdot[\alpha+\beta-1]} \right) \right)^{\cdot[1/\alpha]}, \tag{69}$$

$$\mathbf{A} \leftarrow \mathbf{A} \circledast \left(\left(\mathbf{Z}\, \mathbf{X}^T \right) \oslash \left(\mathbf{Q}^{\cdot[\alpha+\beta-1]}\, \mathbf{X}^T \right) \right)^{\cdot[1/\alpha]}, \tag{70}$$

where $\mathbf{Z} = \mathbf{P}^{\cdot\,[\alpha]} \circledast \mathbf{Q}^{\cdot\,[\beta-1]}$.

In order to fix the scaling indeterminacy between the columns of $\mathbf{A}$ and the rows of $\mathbf{X}$, in practice, after each iteration, we can usually evaluate the l_1-norm of the columns of $\mathbf{A}$ and normalize the elements of the matrices as

$$x_{ij} \leftarrow \quad x_{ij} \textstyle\sum_p a_{pj} \,, \qquad a_{ij} \leftarrow \quad a_{ij} / \displaystyle\sum_p a_{pj} \,. \tag{71}$$

This normalization does not alter $\mathbf{Q} = \mathbf{AX}$, thus, preserving the value of the AB-divergence.

The above novel algorithms are natural extensions of many existing algorithms for NMF, including the ISRA, EMML, Lee-Seung algorithms and Alpha- and Beta-multiplicative NMF algorithms [7,58]. For example, by selecting $\alpha + \beta = 1$, we obtain the Alpha-NMF algorithm, for $\alpha = 1$, we have Beta-NMF algorithms, for $\alpha = -\beta \neq 0$ we obtain a family of multiplicative NMF algorithms based on the extended Itakura-Saito distance [8,53]. Furthermore, for $\alpha = 1$ and $\beta = 1$, we obtain the ISRA algorithm and for $\alpha = 1$ and $\beta = 0$ we obtain the EMML algorithm.

It is important to note that in low-rank approximations, we do not need access to all input data p_{it}. In other words, the above algorithms can be applied for low-rank approximations even if some data are missing or they are purposely omitted or ignored. For large-scale problems, the learning rules can be written in a more efficient form by restricting the generalized mean only to those elements whose indices belong to a preselected subsets $S_T \subset \{1, \ldots, T\}$ and $S_I \subset \{1, \ldots, I\}$ of the whole set of indices.

Using a duality property ($D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = D_{AB}^{(\beta,\alpha)}(\mathbf{Q}\|\mathbf{P})$) of the AB-divergence, we obtain now the dual update rules:

$$\begin{aligned} x_{jt} &\leftarrow x_{jt} \, \exp_{1-\beta}\left(\sum_{i \in S_I} \frac{a_{ij}\, p_{it}^{\lambda-1}}{\sum_{i \in S_I} a_{ij}\, p_{it}^{\lambda-1}} \, \ln_{1-\beta}\left(\frac{q_{it}}{p_{it}}\right)\right), \\ a_{ij} &\leftarrow a_{ij} \, \exp_{1-\beta}\left(\sum_{t \in S_T} \frac{x_{jt}\, p_{it}^{\lambda-1}}{\sum_{t \in S_T} x_{jt}\, p_{it}^{\lambda-1}} \, \ln_{1-\beta}\left(\frac{q_{it}}{p_{it}}\right)\right), \end{aligned} \tag{72}$$

which for $\beta \neq 0$ reduce to

$$\begin{aligned} a_{ij} &\leftarrow a_{ij} \left(\frac{\displaystyle\sum_{t \in S_T} x_{jt}\, p_{it}^{\alpha-1}\, q_{it}^{\beta}}{\displaystyle\sum_{t \in S_T} x_{jt}\, p_{it}^{\alpha+\beta-1}} \right)^{1/\beta}, \\ x_{jt} &\leftarrow x_{jt} \left(\frac{\displaystyle\sum_{i \in S_I} a_{ij}\, p_{it}^{\alpha-1}\, q_{it}^{\beta}}{\displaystyle\sum_{i \in S_I} a_{ij}\, p_{it}^{\alpha+\beta-1}} \right)^{1/\beta}. \end{aligned} \tag{73}$$

3.2. *Conditions for a Monotonic Descent of AB-Divergence*

In this section, we first explain the basic principle of auxiliary function method, which allows us to establish conditions for which NMF update rules provide monotonic descent of the cost function during iterative process. In other words, we analyze the conditions for the existence of auxiliary functions that justify the monotonous descent in the AB-divergence under multiplicative update rules of the form

(61). NMF update formulas with such property have been previously obtained for certain particular divergences: in [38,41] for the Euclidean distance and Kullback-Leibler divergence, in [58] for the Alpha-Divergence and in [59–61] for the Beta-divergence.

Let $\mathbf{Q}^{(k)} = \mathbf{A}^{(k)}\mathbf{X}^{(k)}$ denote our current factorization model of the observations in the k-th iteration, and let $\{\mathbf{Q}^{(k+1)}\}$ denote our candidate model for the next $k+1$-iteration. In the following, we adopt the notation $\{\mathbf{Q}^{(k)}\} \equiv \{\mathbf{A}^{(k)}, \mathbf{X}^{(k)}\}$ to refer compactly to the non-negative factors of the decompositions.

An auxiliary function $G(\{\mathbf{Q}^{(k+1)}\}, \{\mathbf{Q}^{(k)}\}; \mathbf{P})$ for the surrogate optimization of the divergence should satisfy

$$G(\{\mathbf{Q}^{(k+1)}\}, \{\mathbf{Q}^{(k)}\}; \mathbf{P}) \geq G(\{\mathbf{Q}^{(k+1)}\}, \{\mathbf{Q}^{(k+1)}\}; \mathbf{P}) \quad = \quad D_{AB}^{(\alpha,\beta)}(\mathbf{P}||\mathbf{Q}^{(k+1)}), \tag{74}$$

with equality holding for $\{\mathbf{Q}^{(k)}\} = \{\mathbf{Q}^{(k+1)}\}$.

In analogy to the Expectation Maximization techniques developed in [38,41,58], our objective is to find a convex upper bound of the AB divergence and to replace the minimization of the original AB-divergence by an iterative optimization of the auxiliary function.

Given the factors of the current factorization model $\{\mathbf{Q}^{(k)}\}$, the factors of the next candidate $\{\mathbf{Q}^{(k+1)}\}$ are chosen as

$$\{\mathbf{Q}^{(k+1)}\} = \arg\min_{\{\mathbf{Q}\}} G(\{\mathbf{Q}\}, \{\mathbf{Q}^{(k)}\}; \mathbf{P}). \tag{75}$$

where, for simplicity, the minimization is usually carried out with respect to only one of the factors $\mathbf{A}$ or $\mathbf{X}$ of $\{\mathbf{Q}\}$, while keeping the other factor fixed and common in $\{\mathbf{Q}^{(k+1)}\}$, $\{\mathbf{Q}^{(k)}\}$ and $\{\mathbf{Q}\}$.

Assuming that auxiliary function is found, a monotonic descent of the AB-divergence is a consequence of the chain of inequalities:

$$\begin{aligned}
D_{AB}^{(\alpha,\beta)}(\mathbf{P}||\mathbf{Q}^{(k)}) &= G(\{\mathbf{Q}^{(k)}\}, \{\mathbf{Q}^{(k)}\}; \mathbf{P}) && (76)\\
&\overset{(a)}{\geq} G(\{\mathbf{Q}^{(k+1)}\}, \{\mathbf{Q}^{(k)}\}; \mathbf{P}) && (77)\\
&\overset{(b)}{\geq} G(\{\mathbf{Q}^{(k+1)}\}, \{\mathbf{Q}^{(k+1)}\}; \mathbf{P}) && (78)\\
&= D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}^{(k+1)})\,. && (79)
\end{aligned}$$

The inequality (a) is due to the optimization in (75), while the inequality (b) reflects the definition of the auxiliary function in (74).

3.3. A Conditional Auxiliary Function

It is shown in Appendix B that under a certain condition, the function

$$G(\{\mathbf{Q}^{(k+1)}\}, \{\mathbf{Q}^{(k)}\}; \mathbf{P}) = \sum_{i,t}\sum_{j} \gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\})\, d_{AB}^{(\alpha,\beta)}\left(p_{it}, \hat{q}_{it}^{(j)}\right) \tag{80}$$

$$\text{where} \quad \gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\}) = \frac{a_{ij}^{(k)} x_{jt}^{(k)}}{q_{it}^{(k)}} \quad \text{and} \quad \hat{q}_{it}^{(j)} = \frac{a_{ij}^{(k+1)} x_{jt}^{(k+1)}}{\gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\})}, \tag{81}$$

is an auxiliary function for the surrogate optimization of $D_{AB}^{(\alpha,\beta)}(\mathbf{P}||\mathbf{Q})$. The required condition is, at each iteration, the convexity of $d_{AB}^{(\alpha,\beta)}(p_{it}, \hat{q}_{it})$ with respect to all $\hat{q}_{it} \in [\min_j \hat{q}_{it}^{(j)}, \max_j \hat{q}_{it}^{(j)}]$.

Recall that $\mathbf{Q}^{(k)} = \mathbf{A}^{(k)}\mathbf{X}^{(k)}$ corresponds to the current iteration, while a candidate model $\mathbf{Q}^{(k+1)} = \mathbf{A}^{(k+1)}\mathbf{X}^{(k+1)}$ for the next iteration, is the solution of the optimization problem (75). For updating $\mathbf{A}$ at each iteration, we keep $\mathbf{X}^{(k+1)}$ equal to $\mathbf{X}^{(k)}$ and find the global minimum of the auxiliary function $G(\{\mathbf{A}^{(k+1)}, \mathbf{X}^{(k)}\}, \{\mathbf{A}^{(k)}, \mathbf{X}^{(k)}\}; \mathbf{P})$ with respect to $\mathbf{A}^{(k+1)}$.

By setting the following derivative to zero

$$\frac{\partial G(\mathbf{AX}, \mathbf{Q}^{(k)}; \mathbf{P})}{\partial a_{ij}} = a_{ij}^{\beta-1} \sum_{t=1}^{T} (q_{it}^{(k)})^{\lambda-1} x_{jt}^{(k)} \left[\ln_{1-\alpha}\left(\frac{a_{ij}}{a_{ij}^{(k)}}\right) - \ln_{1-\alpha}\left(\frac{p_{it}}{q_{it}^{(k)}}\right)\right] = 0 \,, \tag{82}$$

and solving with respect to $a_{ij} = a_{ij}^{(k+1)}$ yields

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} \exp_{1-\alpha}\left(\sum_{t=1}^{T} \frac{(q_{it}^{(k)})^{\lambda-1} x_{jt}^{(k)}}{\sum_{t=1}^{T} (q_{it}^{(k)})^{\lambda-1} x_{jt}^{(k)}} \ln_{1-\alpha}\left(\frac{p_{it}}{q_{it}^{(k)}}\right)\right) . \tag{83}$$

Analogously, for the intermediate factorized model $\tilde{\mathbf{Q}}^{(k)} = \mathbf{A}^{(k+1)}\mathbf{X}^{(k)}$, the global minimum of the auxiliary function $G(\mathbf{A}^{(k+1)}\mathbf{X}^{(k+1)}, \{\mathbf{A}^{(k+1)}, \mathbf{X}^{(k)}\}; \mathbf{P})$ with respect to the new update $\mathbf{X}^{(k+1)}$, while keeping $\mathbf{A} = \mathbf{A}^{(k+1)}$, is given by

$$x_{jt}^{(k+1)} = x_{jt}^{(k)} \exp_{1-\alpha}\left(\sum_{i=1}^{I} \frac{(\tilde{q}_{it}^{(k)})^{\lambda-1} a_{ij}^{(k+1)}}{\sum_{i=1}^{I} (\tilde{q}_{it}^{(k)})^{\lambda-1} a_{ij}^{(k+1)}} \ln_{1-\alpha}\left(\frac{p_{it}}{\tilde{q}_{it}^{(k)}}\right)\right) . \tag{84}$$

The above two equations match with the updates proposed in (61). As previously indicated, a sufficient condition for a monotonic descent of the AB-divergence is the convexity of terms $d_{AB}^{(\alpha,\beta)}(p_{it}, \hat{q}_{it})$ for all $\hat{q}_{it} \in [\min_j \hat{q}_{it}^{(j)}, \max_j \hat{q}_{it}^{(j)}]$. At this point, it is illustrative to interpret the elements $\hat{q}_{it}^{(j)}$ defined in (81) as linear predictions the factorization model $q_{it} = \sum_j a_{ij}x_{jt}$, obtained from the components $a_{ij}x_{jt}$ and factors $\gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\})$.

Appendix C, provides necessary and sufficient conditions for the convexity of $d_{AB}^{(\alpha,\beta)}(p_{it}, \hat{q}_{it})$ with respect to $\hat{q}_{it}$. Depending on the parameter β, it is required that one of the following condition be satisfied:

$$\begin{cases} \dfrac{p_{it}}{\hat{q}_{it}} \geq c(\alpha, \beta) & \text{for} \quad \beta < \min\{1, 1-\alpha\}, \\ \text{always convex} & \text{for} \quad \beta \in [\min\{1, 1-\alpha\}, \max\{1, 1-\alpha\}], \\ \dfrac{p_{it}}{\hat{q}_{it}} \leq c(\alpha, \beta) & \text{for} \quad \beta > \max\{1, 1-\alpha\}, \end{cases} \tag{85}$$

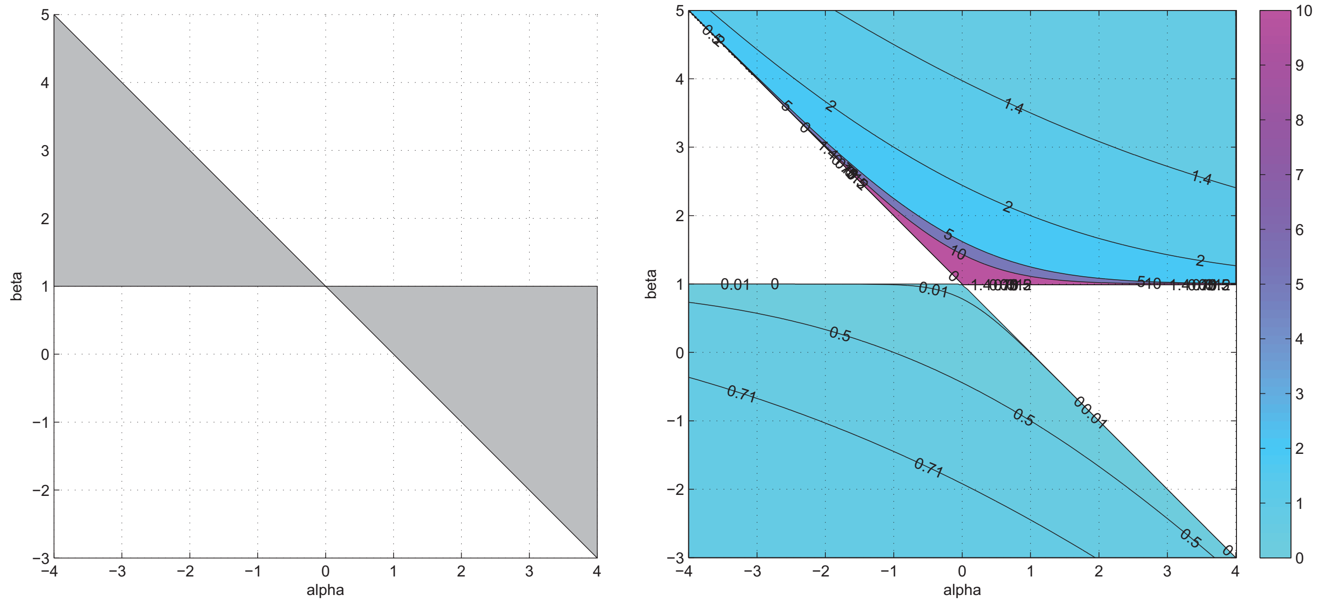
where the upper and lower bounds depend on the function

$$c(\alpha, \beta) = \exp_{1-\alpha}\left(\frac{1}{\beta - 1}\right) , \tag{86}$$

whose contour plot is shown in Figure 4(b).

The convexity of the divergence w.r.t. q_{it} holds for the parameters (α, β) within the convex cone of $\beta \in [\min\{1, 1-\alpha\}, \max\{1, 1-\alpha\}]$. Therefore, for this set of parameters, the monotonic descent in the AB-divergence with the update formulas (83), (84) is guaranteed.

Figure 4. Convexity analysis of $d_{AB}^{(\alpha,\beta)}(p_{it}, \hat{q}_{it})$ w.r.t. $\hat{q}_{it}$ in the (α, β) plane.



(a) The region in gray shows the convex cones, delimited by the lines $\alpha + \beta = 1$ and $\beta = 1$, that guarantee the convexity of $d_{AB}^{(\alpha,\beta)}(\cdot, \cdot)$ with respect to the second argument.

(b) Contour plot of the bound $c(\alpha, \beta)$ on the ratio of $\frac{p_{it}}{\hat{q}_{it}}$. In the plot the areas between isolines are filled using constant colors. The figure reveals a discontinuity towards ∞ in the upper-part and right-part of the straight lines.

On the other hand, even when $\beta \notin [\min\{1, 1-\alpha\}, \max\{1, 1-\alpha\}]$, we can still guarantee the monotonic descent if the set of estimators $\hat{q}_{it} \in [\min_j \hat{q}_{it}^{(j)}, \max_j \hat{q}_{it}^{(j)}]$ is bounded and sufficiently close to p_{it} so that they satisfy the conditions (85). Figure 4(b) illustrates this property demonstrating that if the ratios $\frac{p_{it}}{\hat{q}_{it}}$ approach unity, the convexity of the divergence w.r.t. q_{it} holds for an increasing size of (α, β) plane. In other words, for sufficiently small relative errors between p_{it} and $\hat{q}_{it}$, the update formulas (83), (84) still guarantee a monotonic descent of the AB-divergence, within a reasonable wide range of the hyperparameters (α, β).

3.4. Unconditional Auxiliary Function

The conditions of the previous section for a monotonic descent can be avoided by upper-bounding $d_{AB}^{(\alpha,\beta)}(p_{it}, \hat{q}_{it})$ with another function $\bar{d}_{AB}^{(\alpha,\beta)}\left(p_{it}, \hat{q}_{it}; q_{it}^{(k)}\right)$ which is convex with respect to $\hat{q}_{it}$ and, at the same time, tangent to the former curve at $\hat{q}_{it} = q_{it}^{(k)}$, *i.e.*,

$$d_{AB}^{(\alpha,\beta)}(p_{it}, q_{it}) \leq \bar{d}_{AB}^{(\alpha,\beta)}\left(p_{it}, \hat{q}_{it}; q_{it}^{(k)}\right), \tag{87}$$

where

$$d_{AB}^{(\alpha,\beta)}\left(p_{it}, q_{it}^{(k)}\right) = \bar{d}_{AB}^{(\alpha,\beta)}\left(p_{it}, q_{it}^{(k)}, q_{it}^{(k)}\right), \tag{88}$$

and

$$\frac{\partial^2 \bar{d}_{AB}^{(\alpha,\beta)}\left(p_{it}, \hat{q}_{it}; q_{it}^{(k)}\right)}{\partial \hat{q}_{it}^2} \geq 0. \tag{89}$$

Similarly to the approach in [59–61] for the Beta-divergence, we have constructed an auxiliary function for the AB-divergence by linearizing those additive concave terms of the AB-divergence:

$$\bar{d}_{AB}^{(\alpha,\beta)}\left(p_{it},\hat{q}_{it};q_{it}^{(k)}\right)=\begin{cases}\dfrac{p_{it}^{\alpha+\beta}}{\beta(\alpha+\beta)}+\dfrac{(q_{it}^{(k)})^{\alpha+\beta}}{\alpha(\alpha+\beta)}-\dfrac{p_{it}^{\alpha}\hat{q}_{it}^{\beta}}{\alpha\beta}+\dfrac{(q_{it}^{(k)})^{\alpha+\beta-1}}{\alpha}(\hat{q}_{it}-q_{it}^{(k)}), & \frac{\beta}{\alpha}<\frac{1}{\alpha}-1,\\ \dfrac{p_{it}^{\alpha+\beta}}{\beta(\alpha+\beta)}+\dfrac{\hat{q}_{it}^{\alpha+\beta}}{\alpha(\alpha+\beta)}-\dfrac{p_{it}^{\alpha}\hat{q}_{it}^{\beta}}{\alpha\beta}, & \frac{\beta}{\alpha}\in[\frac{1}{\alpha}-1,\frac{1}{\alpha}],\\ \dfrac{p_{it}^{\alpha+\beta}}{\beta(\alpha+\beta)}+\dfrac{\hat{q}_{it}^{\alpha+\beta}}{\alpha(\alpha+\beta)}-\dfrac{p_{it}^{\alpha}(q_{it}^{(k)})^{\beta}}{\alpha\beta}-\dfrac{p_{it}^{\alpha}(q_{it}^{(k)})^{\beta-1}}{\alpha}(\hat{q}_{it}-q_{it}^{(k)}), & \frac{\beta}{\alpha}>\frac{1}{\alpha}.\end{cases}$$

The upper-bounds for singular cases can be obtained by continuity using L'Hópitals formula.

The unconditional auxiliary function for the surrogate minimization of the AB-divergence $D_{AB}^{(\alpha,\beta)}(\mathbf{P}||\mathbf{Q})$ is now given by

$$\bar{G}(\{\mathbf{Q}\},\{\mathbf{Q}^{(k)}\};\mathbf{P})=\sum_{i,t}\sum_{j}\gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\})\,\bar{d}_{AB}^{(\alpha,\beta)}\left(p_{it},\hat{q}_{it}^{(j)};q_{it}^{(k)}\right),\tag{90}$$

$$\text{where}\quad\gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\})=\frac{a_{ij}^{(k)}x_{jt}^{(k)}}{q_{it}^{(k)}}\quad\text{and}\quad\hat{q}_{it}^{(j)}=\frac{a_{ij}^{(k+1)}x_{jt}^{(k+1)}}{\gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\})}.\tag{91}$$

The alternating minimization of the above auxiliary function with respect to $\mathbf{A}$ and $\mathbf{X}$ yields the following stabilized iterations (with monotonic descent of the AB-divergence):

$$a_{ij}^{(k+1)}=a_{ij}^{(k)}\left[\exp_{1-\alpha}\left(\sum_{t=1}^{T}\frac{(q_{it}^{(k)})^{\lambda-1}x_{jt}^{(k)}}{\sum_{t=1}^{T}(q_{it}^{(k)})^{\lambda-1}x_{jt}^{(k)}}\ln_{1-\alpha}\left(\frac{p_{it}}{q_{it}^{(k)}}\right)\right)\right]^{w(\alpha,\beta)},\tag{92}$$

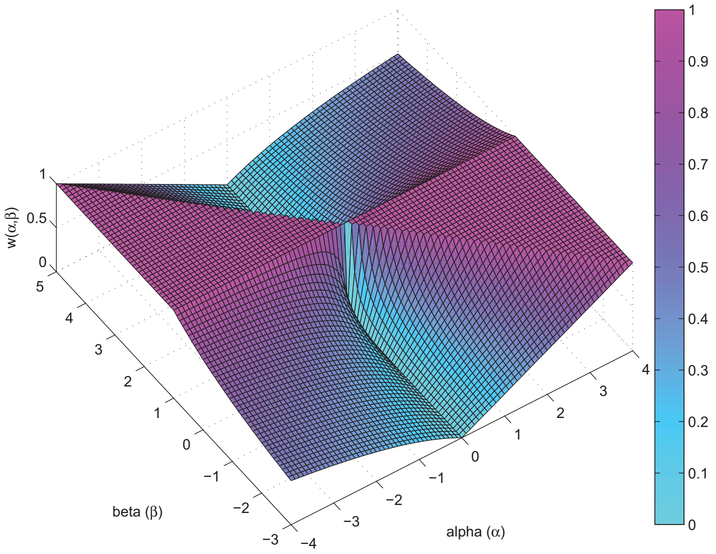
$$x_{jt}^{(k+1)}=x_{jt}^{(k)}\left[\exp_{1-\alpha}\left(\sum_{i=1}^{I}\frac{(\tilde{q}_{it}^{(k)})^{\lambda-1}a_{ij}^{(k+1)}}{\sum_{i=1}^{I}(\tilde{q}_{it}^{(k)})^{\lambda-1}a_{ij}^{(k+1)}}\ln_{1-\alpha}\left(\frac{p_{it}}{\tilde{q}_{it}^{(k)}}\right)\right)\right]^{w(\alpha,\beta)},\tag{93}$$

where

$$w(\alpha,\beta)=\begin{cases}1 & \alpha=0,\ \beta=1,\\ 0 & \alpha=0,\ \beta\neq1,\\ \dfrac{\alpha}{1-\beta} & \alpha\neq0,\ \frac{\beta}{\alpha}<\frac{1}{\alpha}-1,\\ 1 & \alpha\neq0,\ \frac{\beta}{\alpha}\in[\frac{1}{\alpha}-1,\frac{1}{\alpha}],\\ \dfrac{\alpha}{\alpha+\beta-1} & \alpha\neq0,\ \frac{\beta}{\alpha}>\frac{1}{\alpha}.\end{cases}\tag{94}$$

The stabilized formulas (92), (93) coincide with (83), (84), except for the exponent $w(\alpha,\beta)$, which is shown in Figure 5. This exponent is bounded between zero and one, and plays a similar role in the multiplicative update to that of the normalized step-size in an additive gradient descent update. Its purpose is to slow down the convergence so as to guarantee monotonic descent of the cost function. This is a consequence of the fact that the multiplicative correction term is progressively contracted towards the unity as $w(\alpha,\beta)\to0$. For the same reason, the stabilized formulas completely stop the updates for $\alpha=0$ and $\beta\neq1$. Therefore, for α close to zero, we recommend to prevent this undesirable situation by enforcing a positive lower-bound in the value of exponent $w(\alpha,\beta)$.

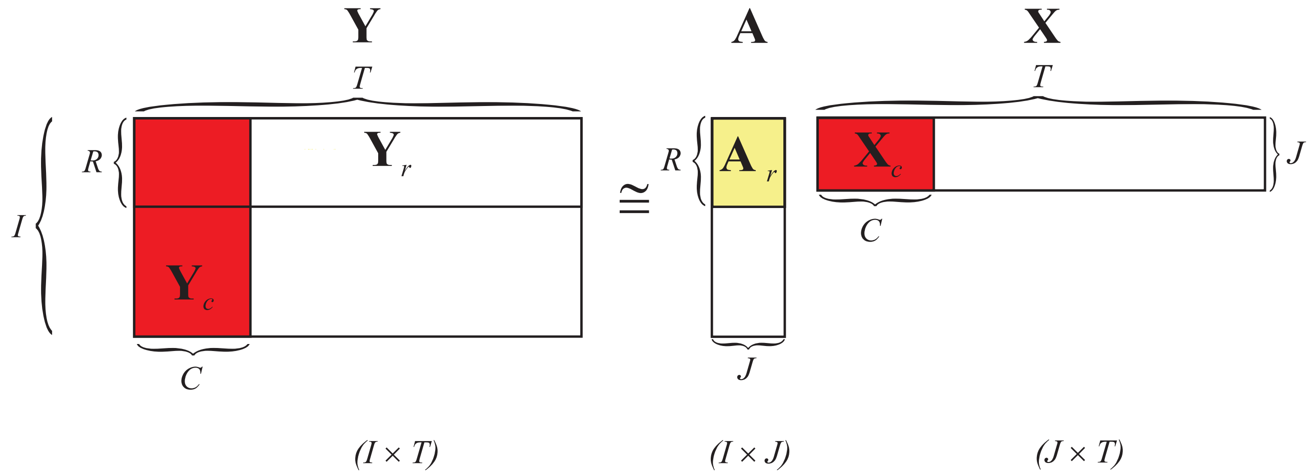
Figure 5. Surface plot of the exponent $w(\alpha, \beta)$, whose role in the multiplicative update is similar to that of a normalized step-size in an additive gradient descent update.



3.5. *Multiplicative NMF Algorithms for Large-Scale Low-Rank Approximation*

In practice, for low-rank approximations with $J \ll \min\{I, T\}$ we do not need to process or save the whole data matrix $\mathbf{Y}$, nor is it necessary to perform computations at each iteration step of the products of the whole estimated matrices $\mathbf{Y}^T\mathbf{A}$ or $\mathbf{Y}\mathbf{X}^T$ (see Figure 6).

Figure 6. Conceptual factorization model of block-wise data processing for a large-scale NMF. Instead of processing the whole matrix $\mathbf{P} = \mathbf{Y} \in \mathbb{R}_+^{I \times T}$, we can process much smaller block matrices $\mathbf{Y}_c \in \mathbb{R}_+^{I \times C}$ and $\mathbf{Y}_r \in \mathbb{R}_+^{R \times T}$ and the corresponding factor matrices $\mathbf{X}_c = \mathbf{B}_c^T = [\boldsymbol{b}_{1,r}, \boldsymbol{b}_{2,r}, \ldots, \boldsymbol{b}_{J,r}]^T \in \mathbb{R}_+^{J \times C}$ and $\mathbf{A}_r = [\boldsymbol{a}_{1,r}, \boldsymbol{a}_{2,r}, \ldots, \boldsymbol{a}_{J,r}] \in \mathbb{R}_+^{R \times J}$ with $J < C << T$ and $J < R << I$. For simplicity of graphical illustration, we have selected the first C columns of the matrices $\mathbf{P}$ and $\mathbf{X}$ and the first R rows of $\mathbf{A}$.



In other words, in order to perform the basic low-rank NMF

$$\mathbf{Y} = \mathbf{P} = \mathbf{A}\mathbf{X} + \mathbf{E},$$

we need to perform two associated nonnegative matrix factorizations using much smaller-scale matrices for large-scale problems, given by

$$\mathbf{Y}_r = \mathbf{A}_r\mathbf{X} + \mathbf{E}_r, \qquad \text{for fixed (known)} \qquad \mathbf{A}_r, \tag{95}$$

$$\mathbf{Y}_c = \mathbf{A}\mathbf{X}_c + \mathbf{E}_c, \qquad \text{for fixed (known)} \qquad \mathbf{X}_c, \tag{96}$$

where $\mathbf{Y}_r \in \mathbb{R}_+^{R\times T}$ and $\mathbf{Y}_c \in \mathbb{R}_+^{I\times C}$ are the matrices constructed from the selected rows and columns of the matrix $\mathbf{Y}$, respectively. Analogously, we can construct the reduced matrices: $\mathbf{A}_r \in \mathbb{R}_+^{R\times J}$ and $\mathbf{X}_c \in \mathbb{R}_+^{J\times C}$ by using the same indices for the columns and rows as those used for the construction of the data sub-matrices $\mathbf{Y}_c$ and $\mathbf{Y}_r$. In practice, it is usually sufficient to choose: $J < R \leq 4J$ and $J < C \leq 4J$.

Using this approach, we can formulate the update learning rule for a large-scale multiplicative NMF as (see Figure 6)

$$\mathbf{X} \;\leftarrow\; \mathbf{X} \circledast \left(\left(\mathbf{A}_r^T(\mathbf{Y}_r^{\cdot\,[\alpha]} \circledast \mathbf{Q}_r^{\cdot\,[\beta-1]})\right) \oslash \left(\mathbf{A}_r^T\mathbf{Q}_r^{\cdot\,[\alpha+\beta-1]}\right)\right)^{\cdot\,[w/\alpha]}, \tag{97}$$

$$\mathbf{A} \;\leftarrow\; \mathbf{A} \circledast \left(\left(\left(\mathbf{Y}_c^{\cdot\,[\alpha]} \circledast \mathbf{Q}_c^{\cdot\,[\beta-1]}\right)\, \mathbf{X}_c^T\right) \oslash \left(\mathbf{Q}_c^{\cdot\,[\alpha+\beta-1]}\, \mathbf{X}_c^T\right)\right)^{\cdot\,[w/\alpha]}, \tag{98}$$

$$\mathbf{Q}_r \;=\; \mathbf{A}_r\,\mathbf{X}, \qquad \mathbf{Q}_c = \mathbf{A}\,\mathbf{X}_c. \tag{99}$$

In fact, we need to save only two reduced set of data matrices $\mathbf{Y}_r \in \mathbb{R}_+^{R\times T}$ and $\mathbf{Y}_c \in \mathbb{R}_+^{I\times C}$. For example, for large dense data matrix $\mathbf{Y} \in \mathbb{R}_+^{I\times T}$ with $I = T = 10^5$ and $J = 10$ and $R = C = 50$, we need to save only 10^7 nonzero entries instead of 10^{10} entries, thus reducing memory requirement 1000 times.

We can modify and extend AB NMF algorithms in several ways. First of all, we can use for update of matrices $\mathbf{A}$ and $\mathbf{X}$ two different cost functions, under the assumption that both the functions are convex with respect to one set of updated parameters (another set is assumed to be fixed). In a special case, we can use two AB-divergences (with two different sets of parameters: one $D_{AB}^{(\alpha_A,\beta_A)}(\mathbf{Y}||\mathbf{A}\mathbf{X})$ for the estimation of $\mathbf{A}$ and fixed $\mathbf{X}$, and another one $D_{AB}^{(\alpha_X,\beta_X)}(\mathbf{Y}||\mathbf{A}\mathbf{X})$ for the estimation of $\mathbf{X}$ and fixed $\mathbf{A}$). This leads to the following updates rules

$$\mathbf{X} \;\leftarrow\; \mathbf{X} \circledast \left(\left(\mathbf{A}_r^T(\mathbf{Y}_r^{\cdot\,[\alpha_X]} \circledast \mathbf{Q}_r^{\cdot\,[\beta_X-1]})\right) \oslash \left(\mathbf{A}_r^T\mathbf{Q}_r^{\cdot\,[\alpha_X+\beta_X-1]}\right)\right)^{\cdot\,[w_X/\alpha_X]}, \tag{100}$$

$$\mathbf{A} \;\leftarrow\; \mathbf{A} \circledast \left(\left(\left(\mathbf{Y}_c^{\cdot\,[\alpha_A]} \circledast \mathbf{Q}_c^{\cdot\,[\beta_A-1]}\right)\, \mathbf{X}_c^T\right) \oslash \left(\mathbf{Q}_c^{\cdot\,[\alpha_A+\beta_A-1]}\, \mathbf{X}_c^T\right)\right)^{\cdot\,[w_A/\alpha_A]}, \tag{101}$$

$$\mathbf{Q}_r \;=\; \mathbf{A}_r\,\mathbf{X}, \qquad \mathbf{Q}_c = \mathbf{A}\,\mathbf{X}_c. \tag{102}$$

In order to accelerate the convergence of the algorithm, we can estimate one of the factor matrices, e.g., $\mathbf{A}$ by using the ALS (Alternating Least Squares) algorithm [7]. This leads to the following modified update rules which can still be robust with respect to $\mathbf{X}$ for suitably chosen set of the parameters.

$$\mathbf{X} \;\leftarrow\; \mathbf{X} \circledast \left(\left(\mathbf{A}_r^T(\mathbf{Y}_r^{\cdot\,[\alpha]} \circledast (\mathbf{A}_r\,\mathbf{X})^{\cdot\,[\beta-1]})\right) \oslash \left(\mathbf{A}_r^T(\mathbf{A}_r\,\mathbf{X})^{\cdot\,[\alpha+\beta-1]}\right)\right)^{\cdot\,[w/\alpha]}, \tag{103}$$

$$\mathbf{A} \;\leftarrow\; \max\{\mathbf{Y}_r\mathbf{X}_r^T(\mathbf{X}_r\mathbf{X}_r^T)^{-1}, \varepsilon\}. \tag{104}$$

Another alternative exists, which allows us to use different cost functions and algorithms for each of the factor matrices.

Furthermore, it would be very interesting to apply AB-multiplicative NMF algorithms for inverse problems in which matrix $\mathbf{A}$ is known and we need to estimate only matrix $\mathbf{X}$ for ill-conditioned and noisy data [62,63].

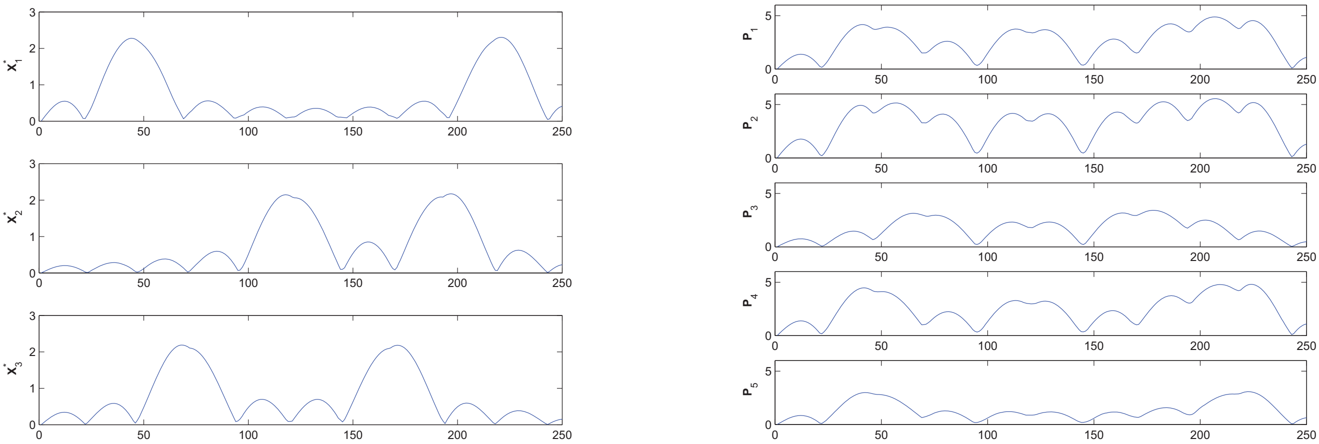
4. Simulations and Experimental Results

We have conducted extensive simulations with experiments designed specifically to address the following aspects:

- What is approximately a range of parameters alpha and beta for which the AB-multiplicative NMF algorithm exhibits the balance between relatively fastest convergence and good performance.
- What is approximately the range of parameters of alpha and beta for which the AB-multiplicative NMF algorithm provides a stable solution independent of how many iterations are needed.
- How robust is the AB-multiplicative NMF algorithm to noisy mixtures under multiplicative Gaussian noise, additive Gaussian noise, spiky biased noise? In other words, find a reasonable range of parameters for which the AB-multiplicative NMF algorithm gives improved performance when the data are contaminated by the different types of noise.

In order to test the performance of the AB-divergence for the NMF problem, we considered the matrix $\mathbf{X}^*$ with three non-negative sources (rows) shown in Figure 7(a). These sources were obtained by superposing pairs of signals from the "ACsincpos10" benchmark of the NMFLAB toolbox and truncating their length to the first 250 samples [7]. We mix these sources with a random mixing matrix $\mathbf{A}^*$ of dimension 25×5, whose elements were drawn independently from a uniform random distribution in the unit interval. This way, we obtain a noiseless observation matrix $\mathbf{P} = \mathbf{A}^*\mathbf{X}^*$ whose first five rows are displayed in Figure 7(b).

Figure 7. Illustration of simulation experiments with three nonnegative sources and their typical mixtures using a randomly generated (uniformly distributed) mixing matrix (rows of a data matrix $\mathbf{P} = \mathbf{AX} + \mathbf{E}$ are denoted by $\mathbf{p}_1, \mathbf{p}_2, ...$).



(a) Three sources (nonnegative components)

(b) Five noiseless mixtures with uniform random mixing matrix

Our objective is to reconstruct from the noisy data $\mathbf{P}$, the matrix $\mathbf{X}$ of the nonnegative sources and the mixing matrix $\mathbf{A}$, by ignoring the scaling and permutation ambiguities.

The performance was evaluated with the mean Signal to Interference Ratio (SIR) of the estimated factorization model $\mathbf{AX}$ and the mean SIR of the estimated sources (the rows of the $\mathbf{X}$) [7].

We evaluate the proposed NMF algorithm in (61) based on the AB-divergence for very large number of pairs of values of (α, β), and a wide range of their values. The algorithm used a single trial random initialization, followed by a refinement which consist of running ten initial iterations of the algorithm with $(\alpha, \beta) = (0.5, 0.5)$. This initialization phase serves to approach the model to the observations ($\mathbf{AX} \to \mathbf{P}$) which is important for guaranteing the posterior monotonic descent in the divergence when the parameters are arbitrary, as discussed in Section 3.3. Then, we ran only 250 iterations of the proposed NMF algorithm for the selected pair of parameters (α, β).

To address the influence of noise, the observations were modeled as $\mathbf{P} = \mathbf{Q}^* + \mathbf{E}$, where $\mathbf{Q}^* = \mathbf{A}^*\mathbf{X}^*$ and $\mathbf{E}$ denote, respectively, the desired components and the additive noise. To cater for different types of noises, we assume that the elements of the additive noise e_{it} are functions of the noiseless model q^*_{it} and of another noise z_{it}, which was independent of q^*_{it}. Moreover, we also assume that the signal q^*_{it} and the noise z_{it} combine additively to give the observations in the deformed logarithm domain as

$$\ln_{1-\alpha^*}(p_{it}) = \ln_{1-\alpha^*}(q^*_{it}) + \ln_{1-\alpha^*}(z_{it}), \tag{105}$$

which is controlled by the parameter α^*. Solving for the observations, we obtain

$$p_{it} = \exp_{1-\alpha^*}\left(\ln_{1-\alpha^*}(q^*_{it}) + \ln_{1-\alpha^*}(z_{it})\right). \tag{106}$$

or equivalently,

$$p_{it} = \begin{cases} q^*_{it}\, z_{it} & , \alpha^* = 0, \quad \text{multiplicative noise,} \\ q^*_{it} + z_{it} & , \alpha^* = 1, \quad \text{additive noise,} \\ \left((q^*_{it})^{\alpha^*} + (z_{it})^{\alpha^*}\right)^{\frac{1}{\alpha^*}} & , \alpha^* \neq 0, \quad \text{additive noise in a deformed log-domain.} \end{cases} \tag{107}$$

Such approach allows us to model, under one single umbrella, noises of multiplicative type, additive noises, and other noises that act additively in a transformed domain, together with their distributions.

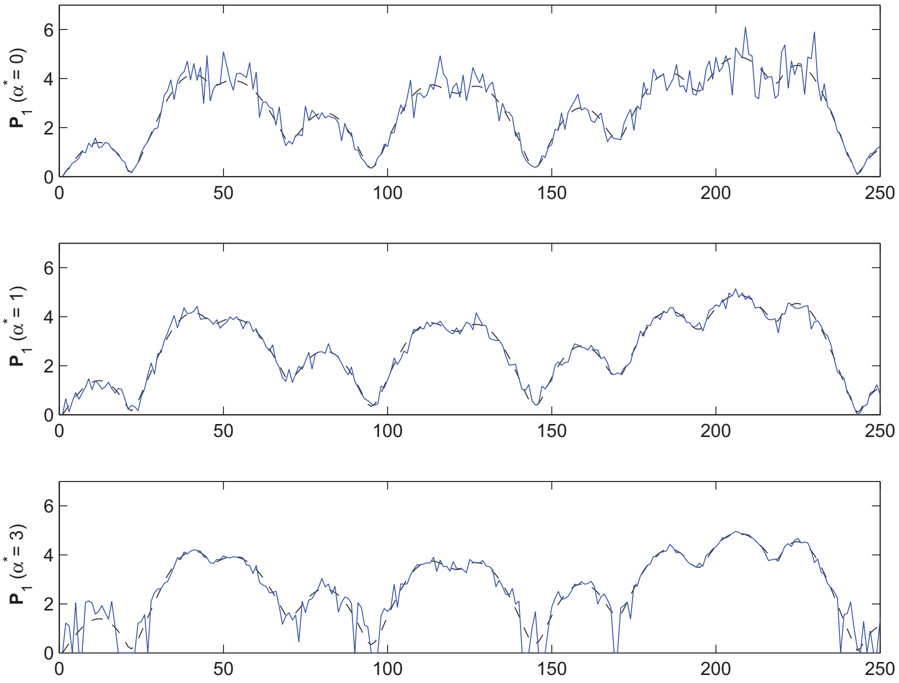
In order to generate the observations, we should assume first a probability density function $g(\bar{z}_{it})$ for $\bar{z}_{it} \equiv \ln_{1-\alpha^*}(z_{it})$, the noise in the transformed domain. This distribution is corrected, when necessary, to the nearby distribution that satisfies the positivity constraint of the observations.

In the following, we assume that the noise in the transformed domain $\bar{z}_{it}$ is Gaussian. Figure 8 presents mixtures obtained for this noise under different values of α^*. The transformation (106) of the Gaussian density $g(\bar{z}_{it})$ leads to the marginal distribution of the observations

$$f(p_{it}|q^*_{it}) = \frac{1}{\sqrt{2\pi}\sigma p_{it}^{1-\alpha^*}}\, e^{-\dfrac{\left(\ln_{1-\alpha^*}(p_{it}) - \ln_{1-\alpha^*}(q^*_{it})\right)^2}{2\sigma^2}}, \quad p_{it} > 0, \tag{108}$$

which can be recognized as a deformed log-Gaussian distribution (of order $1-\alpha^*$), with mean $\ln_{1-\alpha^*}(q^*_{it})$ and variance σ^2. This distribution is exact for multiplicative noise ($\alpha^* = 0$), since in this case no correction of the distribution is necessary to guarantee the non-negativity of the observations. For the remaining cases the distribution is only approximate, but the approximation improves when α^* is not far from zero or when q^*_{it} is sufficient large for all i, t, and for σ^2 sufficient small.

Figure 8. Illustration of the effect of the parameter α^* on the noisy observations ($\mathbf{p}_1$ denotes the first row of the matrix $\mathbf{P}$). Dashed lines corresponds to noiseless mixtures and solid lines to the noisy mixtures that obtained when adding noise in the deformed logarithm ($\ln_{1-\alpha^*}(\cdot)$) domain. The noise distribution was Gaussian of zero mean and with a variance chosen so as to obtain an SNR of 20 dB in the deformed logarithm domain. In the top panel the value of the deformation parameter $\alpha^* = 0$, resulting a multiplicative noise that distorts more strongly signals q^*_{it} with larger values. For the middle panel $\alpha^* = 1$, resulting in an additive Gaussian noise that equally affects all q^*_{it} independently of their values. For the bottom panel, $\alpha^* = 3$, distorting more strongly small values of q^*_{it}.



Interestingly, the log-likelihood of the matrix of observations for mutually independent components, distributed according to (108), is

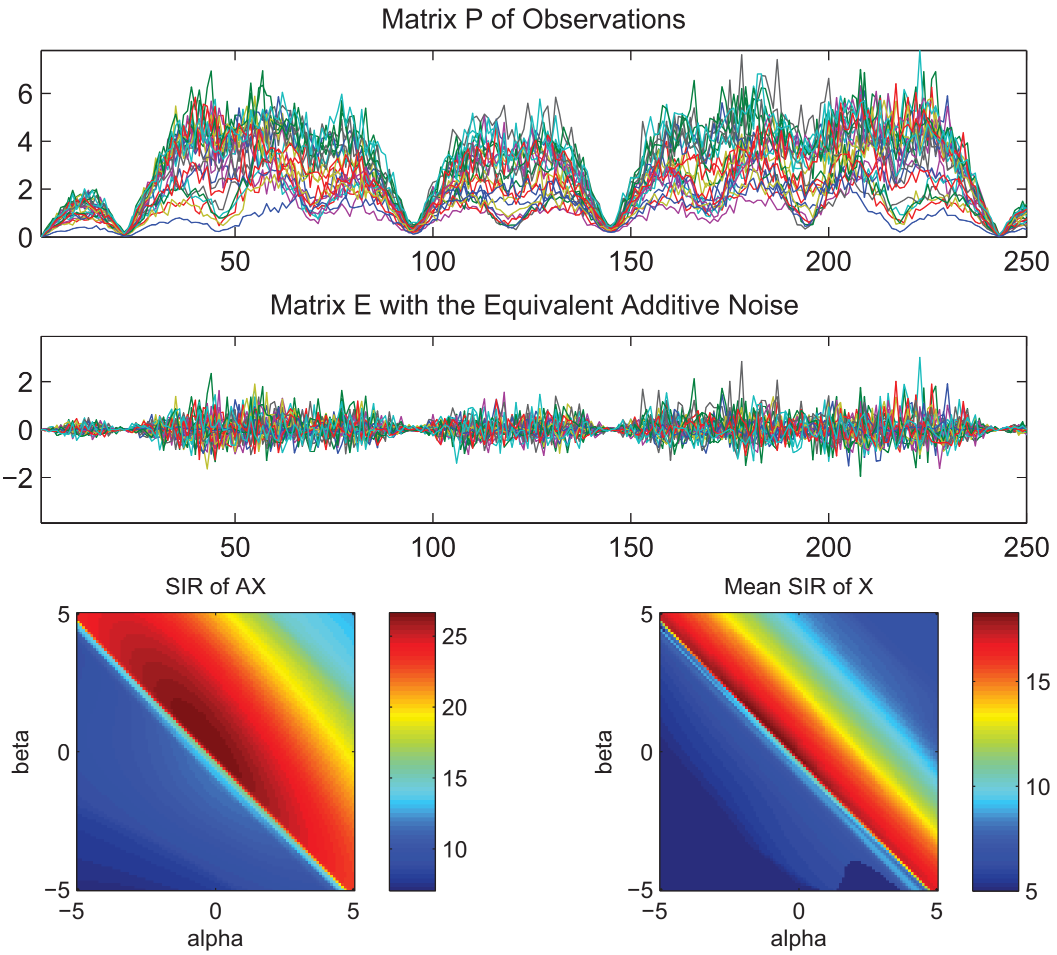
$$\ln f(\mathbf{P}|\mathbf{Q}) \;=\; -\sum_{i,t} \ln\left(\sqrt{2\pi}\sigma p_{it}^{1-\alpha^*}\right) - \frac{1}{\sigma^2} D_{AB}^{(\alpha^*,\alpha^*)}(\mathbf{P}\|\mathbf{Q}) \,. \tag{109}$$

Therefore, provided that (108) is sufficiently accurate, the maximization of the likelihood of the observations is equivalent to the minimization of the AB-divergence for (α^*, α^*), that is,

$$\arg\max_{\mathbf{Q}} \ln f(\mathbf{P}|\mathbf{Q}) \;\equiv\; \arg\min_{\mathbf{Q}} D_{AB}^{(\alpha^*,\alpha^*)}(\mathbf{P}\|\mathbf{Q}) \,. \tag{110}$$

Figure 9 presents the simulation results for mixtures with multiplicative noise. Usually, performances with SIR > 15 dB are considered successful. Observe that the domain where the performance is satisfactory is restricted to $\alpha + \beta \geq -0.5$, otherwise some terms could be extremely small due the inversion of the arguments of the AB-divergence. In other words, for $\alpha + \beta < -0.5$ there is too strong enhancement of the observations that correspond to the values of the model close to zero, deteriorating the performance. In our simulations we restricted the minimum value of entries of the true factorization model to a very small value of 10^{-7}.

Figure 9. Performance of the AB-multiplicative NMF algorithm in the presence of multiplicative noise ($\alpha^* = 0$). The distribution of the noise in the transformed domain $\bar{z}_{it}$ is Gaussian of zero mean and with variance set to obtain an SNR of 20 dB in the $\ln(\cdot)$ domain. The rows of the observation matrix are shown in the top panel, the equivalent additive noise $\mathbf{E} = \mathbf{P} - \mathbf{Q}^*$ is displayed at the middle panel and the performance results are presented at the bottom panels. As theoretically expected, the best SIR of the model (26.7 dB) was achieved in the neighborhood of $(0, 0)$, the parameters for which the likelihood of these observations is maximized. On the other hand, the best mean SIR of the sources (18.0 dB) and of the mixture (21.1 dB) are both obtained for (α, β) close to $(-1.0, 1.0)$.



We confirmed this by extensive computer simulations. As theoretically predicted, for a Gaussian distribution of the noise $\bar{z}_{it}$ with $\alpha^* = 0$, the best performance in the reconstruction of desired components was obtained in the neighborhood of the origin of the (α, β) plane. The generalized Itakura-Saito divergence of Equation (46) gives usually a reasonable or best performance for the estimation of $\mathbf{X}$ and $\mathbf{A}$. The result of the simulation can be interpreted in light of the discussion presented in Section 2.3. As the multiplicative noise increases the distortion at large values of the model, the best performance was obtained for those parameters that prevent the inversion of the arguments of the divergence and, at the same time, suppress the observations that correspond with larger values of the model. This leads to the close to optimal choice of the parameters that satisfy equation $\alpha + \beta = 0$.

The performance for the case of additive Gaussian noise ($\alpha^* = 1$) is presented in Figure 10. For an SNR of 20 dB, the distribution of the noisy observations was approximately Gaussian, since only 1% of the coordinates of the matrix $\mathbf{P}$ were rectified to enforce positivity. This justifies the best performance for the NMF model in the neighborhood of the pair $(\alpha, \beta) = (1, 1)$, since according to (110) minimizing this divergence approximately maximizes the likelihood of these observations. Additionally, we observed poor performance for negative values of α, explained by the large ratios p_{it}/q_{it} being much more unreliable, due to the distortion of the noise for the small values of q^*_{it}, together with the rectifying of the negative observations.

Figure 10. Performance of the AB-multiplicative NMF algorithm for 25 mixtures with additive Gaussian noise and SNR of 20 dB. The best performance for $\mathbf{AX}$ was for an SIR of 31.1 dB, obtained for $(\alpha, \beta) = (0.8, 0.7)$, that is, close to the pair $(1, 1)$ that approximately maximizes the likelihood of the observations. The best performance for $\mathbf{X}$ and $\mathbf{A}$ was obtained in the vicinity of $(-0.2, 0.8)$, with respective mean SIRs of 17.7 dB and 20.5 dB.

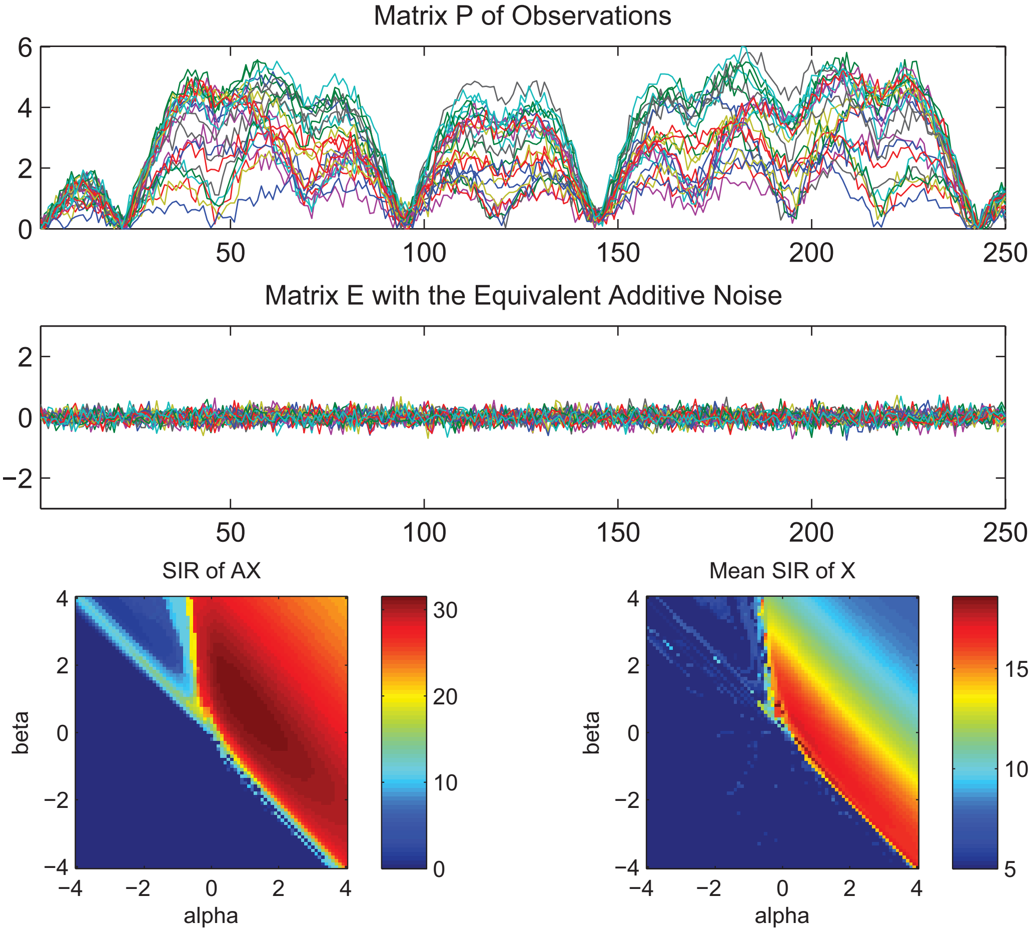
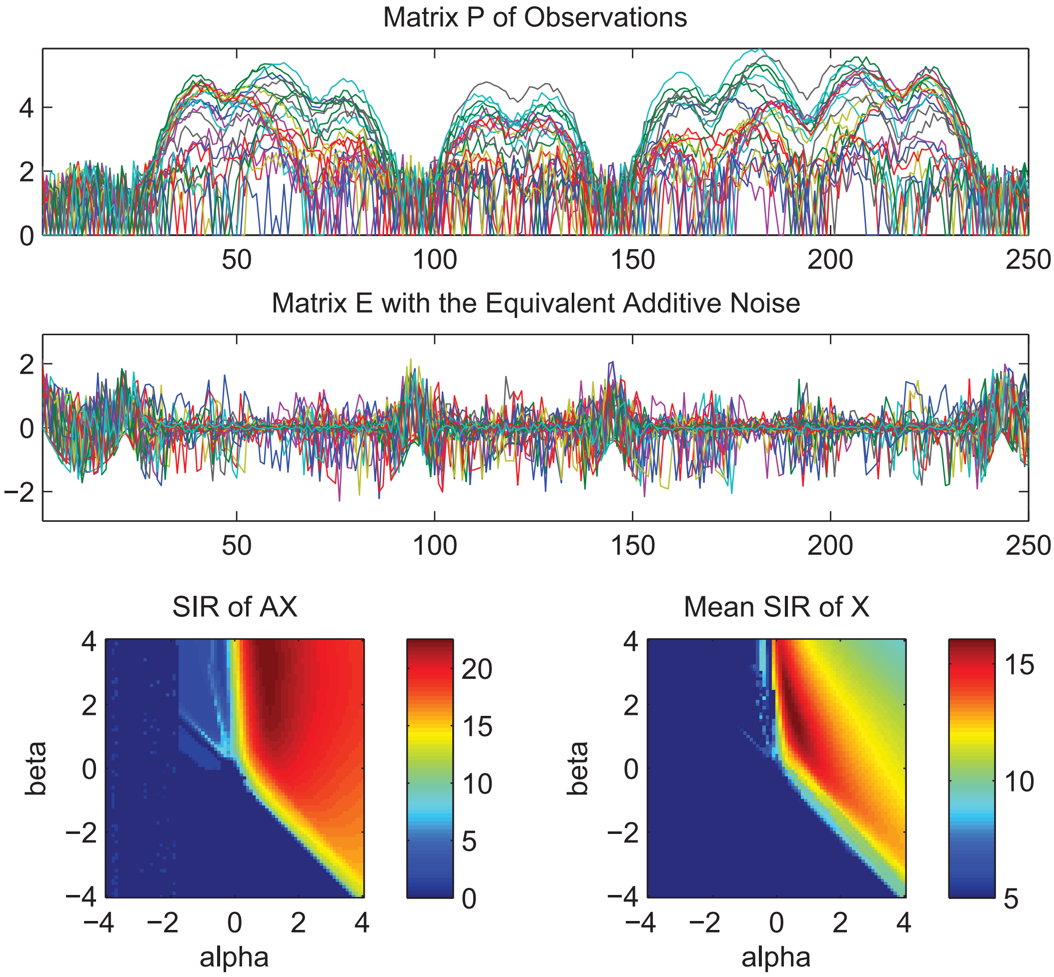


Figure 11 shows simulation results with $\alpha^* = 3$, the case for which the effect of the noise was more pronounced for the small values of q^*_{it}. The noise in the transformed domain was again Gaussian and set to an SNR (in this domain) of 20 dB. In this case, the distribution of the observation was no longer sufficiently close to Gaussian because 11% of the entries of the matrix $\mathbf{P}$ were rectified to enforce their positivity. The graphs reveal that the best performance was obtained for the pair $(\alpha, \beta) = (0.9, 4.0)$ that is quite close to a Beta-divergence. As a consequence of the strong distortion by the noise the small values of the factorization model the large and small ratios p_{it}/q_{it} were not reliable, leaving as preferable the values of α close to unity. On the other hand, the ratios associated with small q_{it} should be penalized, what leads to a larger parameter β as a robust choice against the contamination of the small values of the model.

Figure 11. Performance of the AB-multiplicative NMF algorithm when the observations are contaminated with Gaussian noise in the $\ln_{1-\alpha^*}(\cdot)$ domain, for $\alpha^* = 3$. The best performance for $\mathbf{AX}$ was for an SIR of 22.6 dB obtained for $(\alpha, \beta) = (0.9, 4.0)$. A best SIR of 16.1 dB for $\mathbf{X}$ was obtained for $(\alpha, \beta) = (0.5, 1.7)$, which gave an SIR for $\mathbf{A}$ of 19.1 dB.



The last two simulations, shown in Figures 12 and 13, illustrate the effect of uniform spiky noise which was activated with a probability of 0.1 and contained a bias in its mean. When the observations are biased downwards by the noise the performance improves for a positive α since, as Figure 3 illustrates, in this case the α suppresses the smaller ratios p_{it}/q_{it}. On the other hand, for a positive bias we have the opposite effect, resulting in the negative values of α being preferable. With these changes in α, the value of β should be modified accordingly to be in the vicinity of $-\alpha$ plus a given offset, so as to avoid an excessive penalty for observations that correspond to the large or small values of the factorization model q_{it}.

Figure 12. Performance for biased (non-zero mean) and spiky, additive noise. For $\alpha^* = 1$, we have uniform noise with support in the negative unit interval, which is a spiky or sparse in the sense that it is only activated with a probability of 0.1, *i.e.*, it corrupts only 10% of observed samples. The best SIR results were obtained around the line $(\alpha, 1 - \alpha)$ for both positive and large values of α.

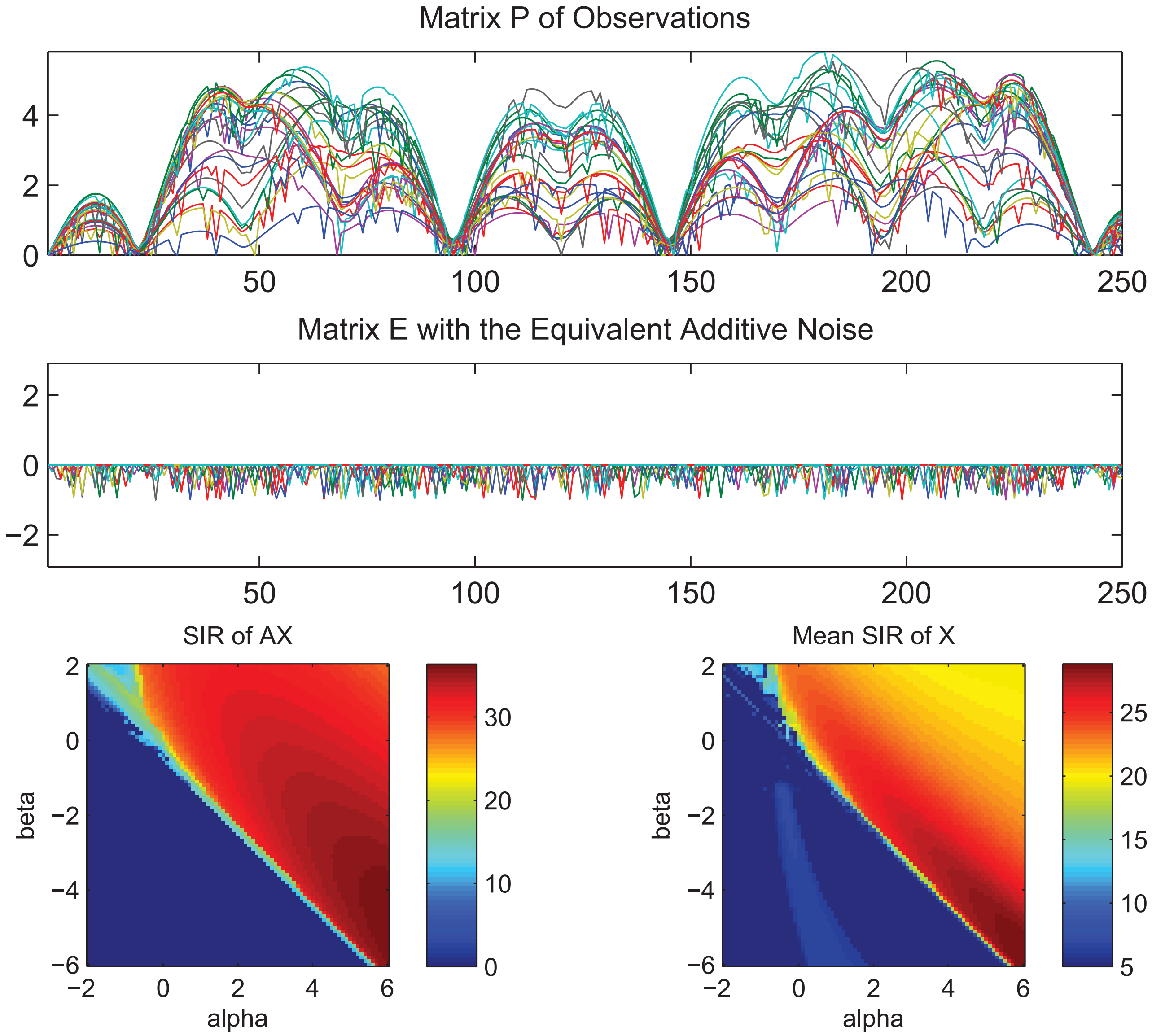
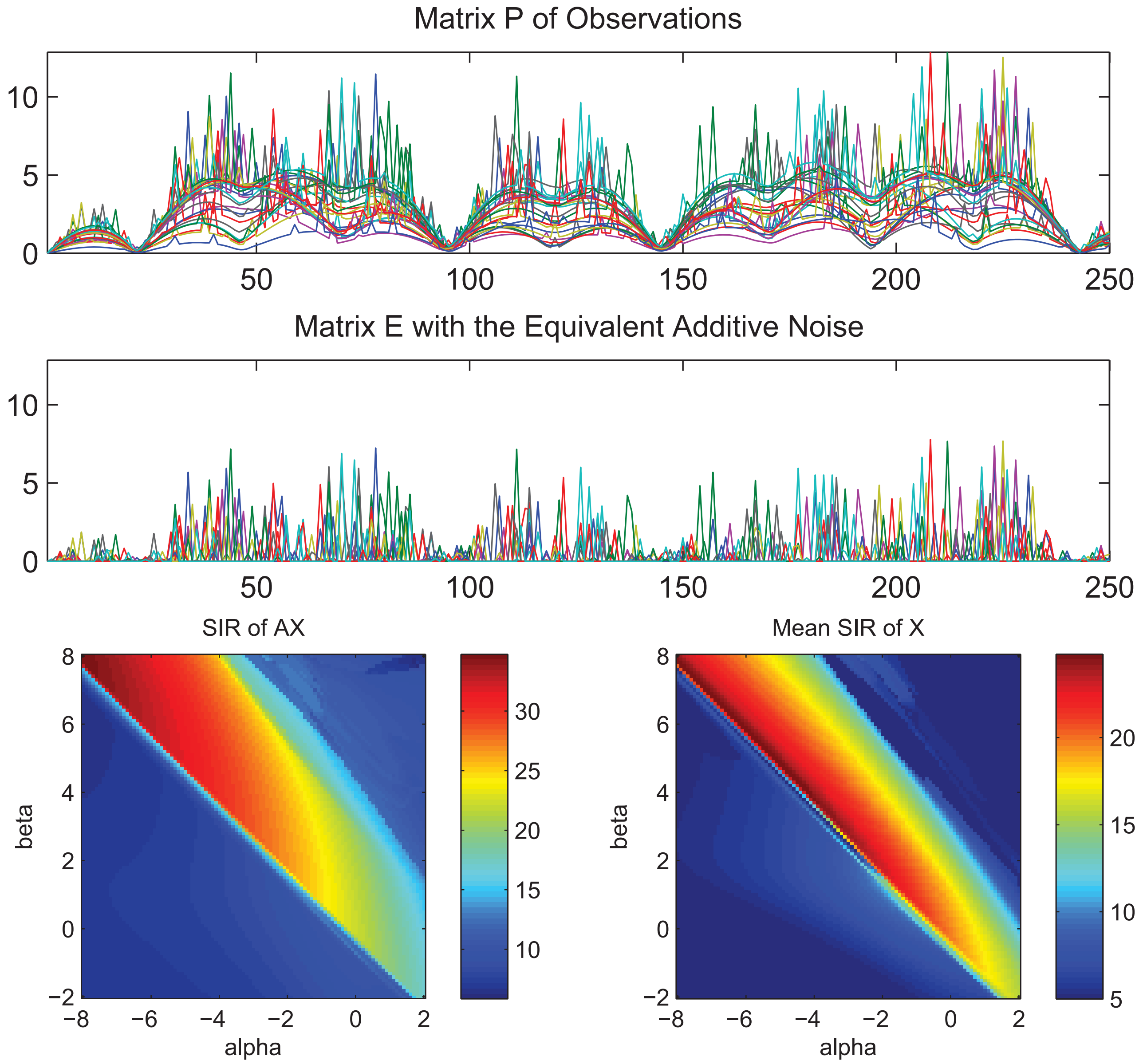


Figure 13. Performance for multiplicative noise that is positively biased and spiky (activated with a probability of 0.1). For $\alpha^* = 0$, the noise in the $\ln(\cdot)$ domain ($\bar{z}_{it}$) followed a uniform distribution with support in the unit interval. The best SIR results were obtained along the line $(\alpha, -\alpha)$ for negative values of α.



5. Conclusions

We have introduced the generalized Alpha-Beta divergence which serves as a flexible and robust cost function and forms a basis for the development of a new class of generalized multiplicative algorithms for NMF. Natural extensions of the Lee-Seung, ISRA, EMML and other NMF algorithms have been presented, in order to obtain more flexible and robust solutions with respect to different contaminations of the data by noise. This class of algorithms allows us to reconstruct (recover) the original signals and to estimate the mixing matrices, even when the observed data are imprecise and/or corrupted by noise.

The optimal choice of the tuning parameters α and β depends strongly both on the distribution of noise and the way it contaminates the data. However, it should be emphasized that we are not expecting that the AB-multiplicative NMF algorithms proposed in this paper will work well for any set of parameters.

In summary, we have rigorously derived a new family of robust AB-multiplicative NMF algorithms using the generalized Alpha-Beta Divergence which unifies and extends a number of the existing divergences. The proposed AB-multiplicative NMF algorithms have been shown to work for wide sets of parameters and combine smoothly many existing NMF algorithms. Moreover, they are also adopted for large-scale low-rank approximation problems. Extensive simulation results have confirmed that the developed algorithms are efficient, robust and stable for a wide range of parameters. The proposed algorithms can also be extended to Nonnegative Tensor Factorizations and Nonnegative Tucker Decompositions in a straightforward manner.

References

1. Amari, S. *Differential-Geometrical Methods in Statistics*; Springer Verlag: New York, NY, USA, 1985.
2. Amari, S. Dualistic geometry of the manifold of higher-order neurons. *Neural Network.* **1991**, *4*, 443–451.
3. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Oxford University Press: New York, NY, USA, 2000.
4. Amari, S. Integration of stochastic models by minimizing α-divergence. *Neural Comput.* **2007**, *19*, 2780–2796.
5. Amari, S. Information geometry and its applications: Convex function and dually flat manifold. In *Emerging Trends in Visual Computing*; Nielsen, F., Ed.; Springer Lecture Notes in Computer Science: Palaiseau, France, 2009a; Volume 5416, pp. 75–102.
6. Amari, S.; Cichocki, A. Information geometry of divergence functions. *Bull. Pol. Acad. Sci. Math.* **2010**, *58*, 183–195.
7. Cichocki, A.; Zdunek, R.; Phan, A.H.; Amari, S. *Nonnegative Matrix and Tensor Factorizations*; John Wiley & Sons Ltd.: Chichester, UK, 2009.
8. Cichocki, A.; Zdunek, R.; Amari, S. Csiszár's divergences for nonnegative matrix factorization: Family of new algorithms. In *Lecture Notes in Computer Science*; Springer: Charleston, SC, USA, 2006; Volume 3889, pp. 32–39.
9. Kompass, R. A Generalized divergence measure for nonnegative matrix factorization. *Neural Comput.* **2006**, *19*, 780–791.
10. Dhillon, I.; Sra, S. Generalized nonnegative matrix approximations with Bregman divergences. *Neural Inform. Process. Syst.* **2005**, 283–290.
11. Amari, S. α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Trans. Inform. Theor.* **2009b**, *55*, 4925–4931.
12. Murata, N.; Takenouchi, T.; Kanamori, T.; Eguchi, S. Convergence-guaranteed multi U-Boost and Bregman divergence. *Neural Comput.* **2004**, *16*, 1437–1481.

13. Fujimoto, Y.; Murata, N. A modified EM Algorithm for mixture models based on Bregman divergence. *Ann. Inst. Stat. Math.* **2007**, *59*, 57–75.
14. Zhu, H.; Rohwer, R. Bayesian invariant measurements of generalization. *Neural Process. Lett.* **1995**, *2*, 28–31.
15. Zhu, H.; Rohwer, R. Measurements of generalisation based on information geometry. In *Mathematics of Neural Networks: Model Algorithms and Applications*; Ellacott, S.W., Mason, J.C., Anderson, I.J., Eds.; Kluwer: Norwell, MA, USA, 1997; pp. 394–398.
16. Nielsen, F.; Nock, R. Sided and symmetrized Bregman centroids. *IEEE Trans. Inform. Theor.* **2009**, *56*, 2882–2903.
17. Boissonnat, J.D.; Nielsen, F.; Nock, R. Bregman Voronoi diagrams. *Discrete Comput. Geom.* **2010**, *44*, 281–307.
18. Yamano, T. A generalization of the Kullback-Leibler divergence and its properties. *J. Math. Phys.* **2009**, *50*, 85–95.
19. Minami, M.; Eguchi, S. Robust blind source separation by Beta-divergence. *Neural Comput.* **2002**, *14*, 1859–1886.
20. Bregman, L. The relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *Comp. Math. Phys. USSR* **1967**, *7*, 200–217.
21. Csiszár, I. Eine Informations Theoretische Ungleichung und ihre Anwendung auf den Beweiss der Ergodizität von Markoffschen Ketten. *Magyar Tud. Akad. Mat. Kutató Int. Közl* **1963**, *8*, 85–108.
22. Csiszár, I. Axiomatic characterizations of information measures. *Entropy* **2008**, *10*, 261–273.
23. Csiszár, I. Information measures: A critial survey. In Proceedings of the Transactions of the 7th Prague Conference, Prague, Czechoslovakia, 18–23 August 1974; pp. 83–86.
24. Ali, M.; Silvey, S. A general class of coefficients of divergence of one distribution from another. *J. Roy. Stat. Soc.* **1966**, *Ser B*, 131–142.
25. Hein, M.; Bousquet, O. Hilbertian metrics and positive definite kernels on probability measures. In Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, Barbados, 6–8 January 2005; pp. 136–143.
26. Zhang, J. Divergence function, duality, and convex analysis. *Neural Comput.* **2004**, *16*, 159–195.
27. Zhang, J.; Matsuzoe, H. Dualistic differential geometry associated with a convex function. In *Springer Series of Advances in Mechanics and Mathematics*,2008; Springer: New York, NY, USA; pp. 58–67.
28. Lafferty, J. Additive models, boosting, and inference for generalized divergences. In Proceedings of the Twelfth Annual Conference on Computational Learning Theory (COLT’99), Santa Cruz, CA, USA, 7-9 July 1999.
29. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
30. Villmann, T.; Haase, S. Divergence based vector quantization using Fréchet derivatives. *Neural Comput.* **2011**, in press.
31. Villmann, T.; Haase, S.; Schleif, F.M.; Hammer, B. Divergence based online learning in vector quantization. In Proceedings of the International Conference on Artifial Intelligence and Soft Computing (ICAISC’2010), LNAI, Zakopane, Poland, 13–17 June 2010.

32. Cichocki, A.; Amari, S.; Zdunek, R.; Kompass, R.; Hori, G.; He, Z. Extended SMART algorithms for nonnegative matrix factorization. In *Lecture Notes in Artificial Intelligence*; Springer: Zakopane, Poland, 2006; Volume 4029, 548–562.
33. Cichocki, A.; Zdunek, R.; Choi, S.; Plemmons, R.; Amari, S. Nonnegative tensor factorization using Alpha and Beta divergences. In IEEE International Conference on Acoustics, Speech, and Signal Processing, Honolulu, Hawaii, USA, 15–20 April 2007; Volume III, pp. 1393–1396.
34. Cichocki, A.; Zdunek, R.; Choi, S.; Plemmons, R.; Amari, S.I. Novel multi-layer nonnegative tensor factorization with sparsity constraints. In *Lecture Notes in Computer Science*; Springer: Warsaw, Poland, 2007; Volume 4432, pp. 271–280.
35. Cichocki, A.; Amari, S. Families of Alpha- Beta- and Gamma- divergences: Flexible and robust measures of similarities. *Entropy* **2010**, *12*, pp. 1532–1568.
36. Paatero, P.; Tapper, U. Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values. *Environmetrics* **1994**, *5*, 111–126.
37. Lee, D.; Seung, H. Learning of the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791.
38. Lee, D.; Seung, H. *Algorithms for Nonnegative Matrix Factorization*; MIT Press: Cambridge MA, USA, 2001; Volume 13, pp. 556–562.
39. Gillis, N.; Glineur, F. Nonnegative factorization and maximum edge biclique problem. ECORE discussion paper 2010. 106 (also CORE DP 2010/59). Available online: http://www.ecore.be/DPs/dp_1288012410.pdf (accessed on 1 November 2008).
40. Daube-Witherspoon, M.; Muehllehner, G. An iterative image space reconstruction algorthm suitable for volume ECT. *IEEE Trans. Med. Imag.* **1986**, *5*, 61–66.
41. De Pierro, A. On the relation between the ISRA and the EM algorithm for positron emission tomography. *IEEE Trans. Med. Imag.* **1993**, *12*, 328–333.
42. De Pierro, A.R. A Modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. *IEEE Trans. Med. Imag.* **1995**, *14*, 132–137.
43. De Pierro, A.; Yamagishi, M.B. Fast iterative methods applied to tomography models with general Gibbs priors. In Proceedings of the SPIE Technical Conference on Mathematical Modeling, Bayesian Estimation and Inverse Problems, Denver, CO, USA, 21 July 1999; Volume 3816, pp. 134–138.
44. Lantéri, H.; Roche, M.; Aime, C. Penalized maximum likelihood image restoration with positivity constraints: multiplicative algorithms. *Inverse Probl.* **2002**, *18*, 1397–1419.
45. Byrne, C. Accelerating the EMML algorithm and related iterative algorithms by rescaled block-iterative (RBI) methods. *IEEE Trans. Med. Imag.* **1998**, *IP-7*, 100–109.
46. Lewitt, R.; Muehllehner, G. Accelerated iterative reconstruction for positron emission tomography based on the EM algorithm for maximum-likelihood estimation. *IEEE Trans. Med. Imag.* **1986**, *MI-5*, 16–22.
47. Byrne, C. *Signal Processing: A Mathematical Approach*. A.K. Peters, Publ.: Wellesley, MA, USA, 2005.
48. Shepp, L.; Vardi, Y. Maximum likelihood reconstruction for emission tomography. *IEEE Trans. Med. Imag.* **1982**, *MI-1*, 113–122.

49. Kaufman, L. Maximum likelihood, least squares, and penalized least squares for PET. *IEEE Trans. Med. Imag.* **1993**, *12*, 200–214.
50. Lantéri, H.; Soummmer, R.; Aime, C. Comparison between ISRA and RLA algorithms: Use of a Wiener filter based stopping criterion. *Astron. Astrophys. Suppl.* **1999**, *140*, 235–246.
51. Cichocki, A.; Zdunek, R.; Amari, S.I. Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization. In *Lecture Notes on Computer Science*; Springer: London, UK, 2007; Volume 4666, pp. 169–176.
52. Minka, T. Divergence measures and message passing. In *Microsoft Research Technical Report*; MSR-TR-2005-173; Microsoft Research Ltd.: Cambridge, UK, 7 December 2005.
53. Févotte, C.; Bertin, N.; Durrieu, J.L. Nonnegative matrix factorization with the Itakura-Saito divergence with application to music analysis. *Neural Comput.* **2009**, *21*, 793–830.
54. Itakura, F.; Saito, F. Analysis synthesis telephony based on the maximum likelihood method. In Proceedings of the 6th International Congress on Acoustics, Tokyo, Japan, 21–28 August 1968; pp. 17–20.
55. Basu, A.; Harris, I.R.; Hjort, N.; Jones, M. Robust and efficient estimation by minimising a density power divergence. *Biometrika* **1998**, *85*, 549–559.
56. Jones, M.; Hjort, N.; Harris, I.R.; Basu, A. A comparison of related density-based minimum divergence estimators. *Biometrika* **1998**, *85*, 865–873.
57. Kivinen, J.; Warmuth, M. Exponentiated gradient versus gradient descent for linear predictors. *Inform. Comput.* **1997**, *132*, 1–63.
58. Cichocki, A.; Lee, H.; Kim, Y.D.; Choi, S. Nonnegative matrix factorization with α-divergence. *Pattern Recogn. Lett.* **2008**, *29*, 1433–1440.
59. Févotte, C. and Idier, J. Algorithms for nonnegative matrix factorization with the β-divergence. Technical Report arXiv **2010**. Available online: http://arxiv.org/abs/1010.1763 (accessed on 13 October 2010).
60. Nakano, M.; Kameoka, H; Le Roux, J; Kitano, Y; Ono, N.; Sagayama, S., Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with β-divergence. In Proceedings of the 2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Kittila, Finland, 29 August–1 September 2010; pp. 283–288.
61. Badeau, R.; Bertin, N.; Vincent, E., Stability analysis of multiplicative updates algorithms and application to nonnegative matirx factroization. *IEEE Trans. Neural Network.* **2010**, 21, 1869–1881.
62. Favati, P.; Lotti, G.; Menchi, O.; Romani, F. Performance analysis of maximum likelihood methods for regularization problems with nonnegativity constraints. *Inverse Probl.* **2010**, *28*. 85013-85030.
63. Benvenuto, F.; Zanella, R.; Zanni, L.; Bertero, M. Nonnegative least-squares image deblurring: Improved gradient projection approaches. *Inverse Probl.* **2010**, *26*, 25004–25021.

Appendix

A. Non-negativity of the AB-divergence

In this appendix we prove that AB-divergence is nonnegative for any values of α and β, and is equal to zero if and only if $\mathbf{P} = \mathbf{Q}$.

For any nonnegative real numbers x and y, and for any positive real numbers a and b that are Hölder conjugate *i.e.*, $a^{-1} + b^{-1} = 1$, Young's inequality states that

$$xy \leq \frac{x^a}{a} + \frac{y^b}{b}, \tag{111}$$

with equality only for $x^a = y^b$.

We shall now show how the non-negativity of the proposed divergence rest on three different inequalities of the Young's type, each one holding true for a different combination of the signs of the constants: $\alpha\beta$, $\alpha(\alpha+\beta)$ and $\beta(\alpha+\beta)$.

For $\alpha\beta > 0$, $\alpha(\alpha+\beta) > 0$ and $\beta(\alpha+\beta) > 0$, we set $x = p_{it}^{\alpha}$, $y = q_{it}^{\beta}$, $a = (\alpha+\beta)/\alpha$ and $b = (\alpha+\beta)/\beta$, to obtain

$$\begin{aligned} \frac{1}{\alpha\beta} p_{it}^{\alpha} q_{it}^{\beta} &\leq \frac{1}{\alpha\beta}\left(\frac{(p_{it}^{\alpha})^{\frac{\alpha+\beta}{\alpha}}}{\frac{(\alpha+\beta)}{\alpha}} + \frac{(q_{it}^{\beta})^{\frac{\alpha+\beta}{\beta}}}{\frac{(\alpha+\beta)}{\beta}}\right) \\ &= \frac{1}{\alpha\beta}\left(\frac{\alpha}{(\alpha+\beta)} p_{it}^{\alpha+\beta} + \frac{\beta}{(\alpha+\beta)} q_{it}^{\alpha+\beta}\right). \end{aligned} \tag{112}$$

For $\alpha\beta < 0$, $\alpha(\alpha+\beta) > 0$ and $\beta(\alpha+\beta) < 0$ we set $x = p_{it}^{\alpha+\beta} q_{it}^{\frac{\beta(\alpha+\beta)}{\alpha}}$, $y = q_{it}^{-\frac{\beta(\alpha+\beta)}{\alpha}}$, $a = \alpha/(\alpha+\beta)$ and $b = -\alpha/\beta$, to obtain

$$\begin{aligned} \frac{-1}{\beta(\alpha+\beta)} p_{it}^{\alpha+\beta} &\leq \frac{-1}{\beta(\alpha+\beta)}\left(\frac{\left(p_{it}^{\alpha+\beta} q_{it}^{\frac{\beta(\alpha+\beta)}{\alpha}}\right)^{\frac{\alpha}{\alpha+\beta}}}{\frac{\alpha}{\alpha+\beta}} + \frac{\left(q_{it}^{-\frac{\beta(\alpha+\beta)}{\alpha}}\right)^{-\frac{\alpha}{\beta}}}{-\frac{\alpha}{\beta}}\right) \\ &= \frac{1}{\alpha\beta}\left(-p_{it}^{\alpha} q_{it}^{\beta} + \frac{\beta}{(\alpha+\beta)} q_{it}^{(\alpha+\beta)}\right). \end{aligned} \tag{113}$$

Finally, for $\alpha\beta < 0$, $\alpha(\alpha+\beta) < 0$ and $\beta(\alpha+\beta) > 0$ we set $x = p_{it}^{\frac{\alpha(\alpha+\beta)}{\beta}} q_{it}^{\alpha+\beta}$, $y = p_{it}^{-\frac{\alpha(\alpha+\beta)}{\beta}}$, $a = \beta/(\alpha+\beta)$ and $b = -\beta/\alpha$, to obtain

$$\begin{aligned} \frac{-1}{\alpha(\alpha+\beta)} q_{it}^{\alpha+\beta} &\leq \frac{-1}{\alpha(\alpha+\beta)}\left(\frac{\left(p_{it}^{\frac{\alpha(\alpha+\beta)}{\beta}} q_{it}^{\alpha+\beta}\right)^{\frac{\beta}{\alpha+\beta}}}{\frac{\beta}{\alpha+\beta}} + \frac{\left(p_{it}^{-\frac{\alpha(\alpha+\beta)}{\beta}}\right)^{-\frac{\beta}{\alpha}}}{-\frac{\beta}{\alpha}}\right) \\ &= \frac{1}{\alpha\beta}\left(-p_{it}^{\alpha} q_{it}^{\beta} + \frac{\alpha}{(\alpha+\beta)} p_{it}^{(\alpha+\beta)}\right). \end{aligned} \tag{114}$$

The three considered cases exhaust all the possibilities for the sign of the constants: $\alpha\beta$, $\alpha(\alpha+\beta)$ and $\beta(\alpha+\beta)$. Inequalities (112)-(114) can be summarized as the joint inequality:

$$\frac{1}{\alpha\beta}p_{it}^{\alpha}q_{it}^{\beta} \leq \frac{1}{\beta(\alpha+\beta)}p_{it}^{\alpha+\beta} + \frac{1}{\alpha(\alpha+\beta)}q_{it}^{\alpha+\beta}, \quad \text{for } \alpha, \beta, \alpha+\beta \neq 0, \tag{115}$$

where the equality holds only for $p_{it} = q_{it}$. This above inequality justifies the non-negativity of the divergence defined in (19).

B. Proof of the Conditional Auxiliary Function Character of $G(\{\mathbf{Q}^{(k+1)}\}, \{\mathbf{Q}^{(k)}\}; \mathbf{P})$

The factorization model of the observations $\mathbf{Q}^{(k)}$ at a k-th iteration is a sum of rank-one matrices, so that

$$q_{it}^{(k)} = \sum_{j=1}^{J} a_{ij}^{(k)} x_{jt}^{(k)} = \sum_{j=1}^{J} \gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\})\, q_{it}^{(k)}, \tag{116}$$

where the parameter

$$\gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\}) = \frac{a_{ij}^{(k)} x_{jt}^{(k)}}{q_{it}^{(k)}} \geq 0, \tag{117}$$

denotes the normalized contribution of the j^{th} rank-one component to the model $q_{it}^{(k)}$. For a given auxiliary function and since $\sum_j \gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\}) = 1$, the elements of the matrix $\mathbf{Q}^{(k+1)}$ resulting from the minimization in (75) can also be expressed by the convex sum

$$q_{it}^{(k+1)} = \sum_{j} \gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\})\, \hat{q}_{it}^{(j)} \quad \text{where} \quad \hat{q}_{it}^{(j)} = \frac{a_{ij}^{(k+1)} x_{jt}^{(k+1)}}{\gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\})}, \qquad j = 1, 2, \ldots, J. \tag{118}$$

In this context the $\hat{q}_{it}^{(j)}$ elements represent a prediction of $q_{it}^{(k+1)}$ obtained from the j^{th}-rank-one component $(a_{ij}^{(k+1)} x_{jt}^{(k+1)})$, assuming that proportions of the components in the model are still governed by $\gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\})$.

If, additionally, the convexity of $d_{AB}^{(\alpha,\beta)}(p_{it}, \hat{q}_{it})$ w.r.t. the second argument holds true in the interval $[\min_j \hat{q}_{it}^{(j)}, \max_j \hat{q}_{it}^{(j)}]$ and for each i, t, the function

$$G(\{\mathbf{Q}^{(k+1)}\}, \{\mathbf{Q}^{(k)}\}; \mathbf{P}) = \sum_{i,t}\sum_{j} \gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\})\, d_{AB}^{(\alpha,\beta)}\left(p_{it}, \hat{q}_{it}^{(j)}\right), \tag{119}$$

can be lower-bounded by means of Jensen's inequality

$$\sum_{i,t}\sum_{j} \gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\})\, d_{AB}^{(\alpha,\beta)}\left(p_{it}, \hat{q}_{it}^{(j)}\right) \geq \sum_{i,t} d_{AB}^{(\alpha,\beta)}\left(p_{it}, \sum_{j} \gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\})\, \hat{q}_{it}^{(j)}\right). \tag{120}$$

>From (118), observe that the previous lower bound is the AB-divergence

$$\sum_{i,t} d_{AB}^{(\alpha,\beta)}\left(p_{it}, \sum_{j} \gamma_{it}^{(j)}(\{\mathbf{Q}^{(k)}\})\, \hat{q}_{it}^{(j)}\right) = D_{AB}^{(\alpha,\beta)}\left(\mathbf{P}||\mathbf{Q}^{(k+1)}\right). \tag{121}$$

thus confirming that

$$G(\{\mathbf{Q}^{(k+1)}\},\{\mathbf{Q}^{(k)}\};\mathbf{P}) \geq D_{AB}^{(\alpha,\beta)}(\mathbf{P}||\mathbf{Q}^{(k+1)}) = G(\{\mathbf{Q}^{(k+1)}\},\{\mathbf{Q}^{(k+1)}\};\mathbf{P}), \tag{122}$$

and proving the desired result: $G(\{\mathbf{Q}^{(k+1)}\},\{\mathbf{Q}^{(k)}\};\mathbf{P})$ is an auxiliary function for the surrogate optimization of the AB-divergence provided that the convexity of $d_{AB}^{(\alpha,\beta)}(p_{it},\hat{q}_{it})$ for $\hat{q}_{it} \in [\min_j \hat{q}_{it}^{(j)}, \max_j \hat{q}_{it}^{(j)}]$ and each i,t, is guaranteed at each iteration.

C. Necessary and Sufficient Conditions for Convexity

The convexity of the divergence with respect to the model (the second argument of the divergence) is an important property in the designing of update formulas with monotonic descent. The second order partial derivative of the divergence in (23) with respect to the second argument is given by

$$\frac{\partial^2 d_{AB}^{(\alpha,\beta)}(p_{it},q_{it})}{\partial q_{it}^2} = \left(1 + (1-\beta)\ln_{1-\alpha}\left(\frac{p_{it}}{q_{it}}\right)\right) q_{it}^{\alpha+\beta-2}. \tag{123}$$

For $\alpha = 0$, the proof of convexity for $\beta = 1$ is obvious from Equation (123). On the other hand, for $\alpha \neq 0$, after the substitution of the deformed logarithm in (123) by its definition, a sufficient condition for the nonnegativity of the second partial derivative

$$\begin{aligned} \frac{\partial^2 d_{AB}^{(\alpha,\beta)}(p_{it},q_{it})}{\partial q_{it}^2} &= \left(1 + \frac{1-\beta}{\alpha}\left[\left(\frac{p_{it}}{q_{it}}\right)^{\alpha} - 1\right]\right) q_{it}^{\alpha+\beta-2}, & (124)\\ &\geq \left(1 - \frac{1-\beta}{\alpha}\right) q_{it}^{\alpha+\beta-2}, & (125)\\ &\geq 0, & (126) \end{aligned}$$

is given by

$$\frac{1-\beta}{\alpha} \in [0,1]. \tag{127}$$

After combining the cases in a single expression the sufficient condition for the divergence $d_{AB}^{(\alpha,\beta)}(p_{it},q_{it})$ to be convex w.r.t. the second argument becomes

$$\beta \in [\min\{1, 1-\alpha\}, \max\{1, 1-\alpha\}]. \tag{128}$$

Figure 4(a) illustrates the domain where this sufficient condition is satisfied in the (α,β) plane.

We can continue further with the analysis to obtain the necessary and sufficient conditions for the convexity after solving directly (123). Depending on the value of β, one of the following conditions should be satisfied

$$\begin{cases} \dfrac{p_{it}}{q_{it}} \geq c(\alpha,\beta) & \text{for} \quad \beta < \min\{1, 1-\alpha\}, \\ \text{always convex} & \text{for} \quad \beta \in [\min\{1, 1-\alpha\}, \max\{1, 1-\alpha\}], \\ \dfrac{p_{it}}{q_{it}} \leq c(\alpha,\beta) & \text{for} \quad \beta > \max\{1, 1-\alpha\}, \end{cases} \tag{129}$$

where the upper and lower bounds depend on the function

$$c(\alpha,\beta) = \exp_{1-\alpha}\left(\frac{1}{\beta-1}\right). \tag{130}$$

A contour plot of $c(\alpha, \beta)$ is shown in Figure 4(b).