Numerical Algorithms
https://doi.org/10.1007/s11075-024-01886-w

ORIGINAL PAPER

On solving a revised model of the nonnegative matrix factorization problem by the modified adaptive versions of the Dai–Liao method

Saman Babaie-Kafaki[1] · Fatemeh Dargahi[2] · Zohre Aminifard[3]

Received: 13 October 2023 / Accepted: 10 July 2024
© The Author(s) 2024

Abstract
We suggest a revised form of a classic measure function to be employed in the optimization model of the nonnegative matrix factorization problem. More exactly, using sparse matrix approximations, the revision term is embedded to the model for penalizing the ill-conditioning in the computational trajectory to obtain the factorization elements. Then, as an extension of the Euclidean norm, we employ the ellipsoid norm to gain adaptive formulas for the Dai–Liao parameter in a least-squares framework. In essence, the parametric choices here are obtained by pushing the Dai–Liao direction to the direction of a well-functioning three-term conjugate gradient algorithm. In our scheme, the well-known BFGS and DFP quasi–Newton updating formulas are used to characterize the positive definite matrix factor of the ellipsoid norm. To see at what level our model revisions as well as our algorithmic modifications are effective, we seek some numerical evidence by conducting classic computational tests and assessing the outputs as well. As reported, the results weigh enough value on our analytical efforts.

Keywords Unconstrained optimization · Nonnegative matrix factorization · Conjugate gradient algorithm · Quasi–Newton updating formula · Ellipsoid norm

Mathematics Subject Classification (2010) 65K05 · 15A23 · 90C53

✉ Saman Babaie-Kafaki
saman.babaiekafaki@unibz.it

1 Faculty of Engineering, Free University of Bozen-Bolzano, Piazza Università 5, 39100 Bolzano, Italy

2 Department of Mathematics, Semnan University, Semnan, Iran

3 UCLouvain, Institute of Information and Communication Technologies, Electronics and Applied Mathematics, Louvain-la-Neuve, Belgium

Published online: 24 July 2024

1 Introduction

A cursory readout of the literature confirms that high-dimensional models have increasingly appeared in the data mining procedures, in the current age of social networks, bioinformatics, digital communications, and quantum computing. This fact places great importance on the necessity of diversifying the strategies for managing the difficulties that need to be prevailed when working with the complex, massive data sets.

A well-known plan to handle the high-dimensional models has been mainly centered on the compact representation of the input data sets [16]. In this regard, data reduction principally targets decreasing the size of the data sets while maintaining the important information, sometimes by data encoding procedures [20]. Meanwhile, when the data sets are given in the matrix forms, classic tools of the linear algebra such as nonnegative matrix factorization (NMF) may be greatly and influentially helpful [10, 17]. As known, a wide range of the real-world data sets are inherently nonnegative and so, we should technically try to rule out the generation of the negative entries while managing and processing such data. Nowadays, NMF is repeatedly and purposefully applied in practical studies such as pattern recognition [11], recommendation systems [21] and face detection [29].

In a common framework, various NMF techniques take a matrix with nonnegative entries as the input, and deliver two lower dimension matrices with nonnegative entries as the output [16], in a way that multiplying the output matrices yields an accurate approximation for the input matrix. As a matter of fact, well-conditioning the intermediary consecutive approximations of the factorization elements may influentially enhance the computational stability [30], and as a result, make it possible to gain more appropriate output matrices as well.

Researchers have recently also pushed to devise memoryless versions of the classic algorithms as another move to handle the high-dimensional optimization models. To contrive a memoryless technique for a general minimization model, we should tactfully benefit the differential features of the cost function as well as the constraints. Meanwhile, the algorithmic steps should be simply performed, not being so time-consuming and labor-intensive, alongside keeping the accuracy at an acceptable level and ensuring the convergence of the solution trajectory. These features can be aggregately seen in the conjugate gradient (CG) algorithms which have been traditionally shaped in the vector forms [28]. Especially, the Dai–Liao (DL) method is nowadays labeled as an efficient CG algorithm due to flexibly incorporating the conjugacy and the quasi–Newton aspects in general circumstances [8, 13].

Here, we plan to address possible model revisions as well as algorithmic modifications of some classic strategies for managing the large-scale data sets. To summarize the organization of our study, firstly we deal with a revised form of the classic measure function proposed by Dennis and Wolkowicz [14] in Section 2, to be embedded to the optimization model of the NMF problem, by penalizing the ill-conditioned intermediary approximations of the factorization elements. Then, in Section 3, we focus on determining adaptive formulas for the DL parameter as the solutions of a least-squares model formulated based on the ellipsoid vector norm [28]. We carry out numerical tests to mirror the value of our theoretical efforts in Section 4, on the CUTEr problems

[18] as well as a set of randomly generated NMF cases. Finally, we summarize some results for better understanding of the progress level in Section 5.

2 A revised model for the nonnegative matrix factorization problem

Dimensionality reduction methodologies are naturally understood as influential approaches for analyzing large data sets. As known, high-dimensional data analysis is an integral part of the digital era due to recent developments in sensor technology. As mentioned in Section 1, NMF is one such techniques that has caught researchers' imagination thanks to the interpretability, simplicity, flexibility and generality [11, 21, 24, 27, 29].

Extracting hidden and important features from data gives rise to the NMF popularity in which the data matrix is approximated by the product of two matrices, usually much smaller than the original data matrix. All the input and output matrices of NMF (often) should be component wisely nonnegative. Mathematically speaking, for a given component wisely nonnegative matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ (or $\mathbf{A} \geq 0$ for short) and a positive integer $r \ll \min\{m, n\}$, NMF entails finding component wisely nonnegative matrices $\mathbf{W} \in \mathbb{R}^{m\times r}$ and $\mathbf{Z} \in \mathbb{R}^{r\times n}$ (or $\mathbf{W} \geq 0$ and $\mathbf{Z} \geq 0$ for short), by solving the following minimization problem:

$$\min_{\mathbf{W}\geq 0,\ \mathbf{Z}\geq 0} \mathfrak{F}(\mathbf{W}, \mathbf{Z}) = \frac{1}{2}\|\mathbf{A} - \mathbf{W}\mathbf{Z}\|_F^2, \tag{2.1}$$

where $\|.\|_F$ stands for the Frobenius norm. In an efficient approach to address (2.1), the alternating nonnegative least-squares (ANLS) technique targets the following two subproblems [22]:

$$\mathbf{Z}^{k+1} = \arg\min_{\mathbf{Z}\geq 0} \mathfrak{F}(\mathbf{W}^k, \mathbf{Z}), \tag{2.2}$$

$$\mathbf{W}^{k+1} = \arg\min_{\mathbf{W}\geq 0} \mathfrak{F}(\mathbf{W}, \mathbf{Z}^{k+1}), \tag{2.3}$$

for all $k \in \mathbb{Z}^+ = \mathbb{N}\bigcup\{0\} = \{0, 1, 2, \dots\}$.

As known, in the computational and analytical studies of the matrix spaces, a great deal of concern is devoted to the matrix condition number, an influential factor that is in a straight connection with the collinearity between the rows or the columns of the matrix [30]. Experiential efforts of the literature show that ill-conditioning may significantly deflect the solution process and yield misleading results. So, it is a classic matter of routine to devise a plan for having control over the condition number of the matrices that iteratively generate in an algorithmic procedure.

A cursory glimpse of the NMF literature shows a lack of analytical will as well as structural tendency to dealing with well-conditioning of the NMF outputs. It should be noted that various modified NMF models mainly target the orthogonality or symmetrization of the decomposition elements [17], being helpful in special applications of the data mining such as sparse recovery and clustering. Such extensions of the classic NMF model have been devised by imposing extra constraints to push the solution

path toward the desired outputs. As a results, the solution process of the mentioned models can be to some extent challenging and sometimes, the workload may get heavy.

To depict the effect of ill-conditioning on the NMF model, here we report the outputs of the MATLAB function 'nnmf' on the well-known Hilbert matrix. Defined by

$$\mathcal{H}_{ij} = \frac{1}{i+j-1},\ i, j = 1, 2, ..., n,$$

the Hilbert matrix $\mathcal{H} \in \mathbb{R}^{n\times n}$ has been classically recognized as an ill-conditioned matrix, being also (symmetric) positive definite. By setting $n = 20$ and $r = 6$, and then investigating the NMF outputs on $\mathcal{H}$ obtained by 10000 different implementations of the MATLAB function 'nnmf', we observed that for more than 46% of the implementations, at least three columns (and rows) of $\mathbf{W}$ (and $\mathbf{Z}$) were equal to zero. That means for more than 46% of the implementations the outputs for $r = 4, 5, 6$ were quite the same. So, in such situations, the NMF cannot serve as a reliable tool in a recommender system for which filling the zero entries (empty positions) is of great importance. On the other hand, we observed that for at least 34% of the outputs the relative error was more than one. These observations could motivate us to deal with collinearity in the NMF model.

Combating the collinearity between the columns of $\mathbf{W}$ or the rows of $\mathbf{Z}$, in order to take computational stability attitude toward the NMF model prompted us to plug condition number of the matrices $\mathcal{W} = \mathbf{W}^T\mathbf{W}$ and $\mathcal{Z} = \mathbf{Z}\mathbf{Z}^T$ of the dimension $r \times r$ into the model (2.1). Note that the existence of sufficient (numerical) linear independency between the columns of $\mathbf{W}$ or the rows of $\mathbf{Z}$, makes the matrices $\mathcal{W}$ and $\mathcal{Z}$ acceptably well-conditioned and positive definite. While, the mentioned collinearity pushes $\mathcal{W}$ and $\mathcal{Z}$ toward ill-conditioning and positive semidefiniteness. So, to be cautious about such troubling issues, the following revised version of the NMF model (2.1) can be proffered:

$$\begin{aligned}\hat{\mathfrak{F}}(\mathbf{W}, \mathbf{Z}) &= \frac{1}{2}\|\mathbf{A} - \mathbf{W}\mathbf{Z}\|_F^2 + \lambda_1\kappa(\mathbf{W}^T\mathbf{W}) + \lambda_2\kappa(\mathbf{Z}\mathbf{Z}^T) \\ &= \frac{1}{2}\|\mathbf{A} - \mathbf{W}\mathbf{Z}\|_F^2 + \lambda_1\kappa(\mathcal{W}) + \lambda_2\kappa(\mathcal{Z}) \\ &= \frac{1}{2}\|\mathbf{A} - \mathbf{W}\mathbf{Z}\|_F^2 + \lambda_1\frac{\text{maxmag}(\mathcal{W})}{\text{minmag}(\mathcal{W})} + \lambda_2\frac{\text{maxmag}(\mathcal{Z})}{\text{minmag}(\mathcal{Z})},\end{aligned} \tag{2.4}$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are the penalty parameters [25, 26] and the maximum magnification (maxmag) and the minimum magnification (minmag) by an arbitrary matrix $P \in \mathbb{R}^{m\times n}$ are respectively defined in Watkins [30] as

$$\text{maxmag}(P) = \max_{x\neq 0}\frac{\|Px\|}{\|x\|},\ \text{and}\ \text{minmag}(P) = \min_{x\neq 0}\frac{\|Px\|}{\|x\|}.$$

As seen, ill-conditioned choices for $\mathbf{W}$ and $\mathbf{Z}$ meaningfully impose penalty to the model. Meanwhile, although seldom occurs in practice, $\hat{\mathfrak{F}}(\mathbf{W}, \mathbf{Z})$ is not well-defined when $\mathbf{W}$ or $\mathbf{Z}$ are rank deficient.

In the model (2.4) well-conditioning has been brought up by straightly embedding penalty terms to the cost function. So, in this respect, since we made the solution process away from the possible troubling consequences resulted by imposing an extra set of constraints, finding approximate solutions of the model may be less challenging. However, we should not overlook the complexity of doing computations by the spectral condition number in the model, especially in large-scale cases. It is generally a matter of fact that carrying out calculations with high-dimensional dense matrices causes extra CPU time and may increase the numerical errors as well. So, developing sparse approximations of such matrices in the data analysis has recently attracted special attentions [25, 26].

Among the fundamental sparse structures for the symmetric matrices, there exist the diagonal and the (banded) symmetric tridiagonal matrices [7] as well as the symmetric rank-one or rank-two updates of the (scaled) identity matrix [28]. In essence, we should conduct a cost-benefit analysis to select a special sparse matrix structure which is of enough suitability in the relevant application. Driven by this issue, because of the presence of the spectral condition number in the augmented model (2.4) which is directly linked to the eigenvalues of the matrix, to tackle some precarious situations stemming from a great deal of time-consuming for calculating $\mathcal{W}$ and $\mathcal{Z}$, it may be preferable to use diagonal approximations of $\mathcal{W}$ and $\mathcal{Z}$ in the model (2.4) by

$$\mathcal{W} \approx \hat{\mathbf{D}} = \text{diag}(\hat{\mathbf{D}}_1^*, \hat{\mathbf{D}}_2^*, ..., \hat{\mathbf{D}}_r^*),$$
$$\mathcal{Z} \approx \check{\mathbf{D}} = \text{diag}(\check{\mathbf{D}}_1^*, \check{\mathbf{D}}_2^*, ..., \check{\mathbf{D}}_r^*),$$

where

$$\hat{\mathbf{D}}_j^* = \sum_{i=1}^{m} \mathbf{W}_{ij}^2, \ j = 1, 2, ..., r, \text{ and } \check{\mathbf{D}}_i^* = \sum_{j=1}^{n} \mathbf{Z}_{ij}^2, \ i = 1, 2, ..., r.$$

Notably, the above diagonal estimations are derived from

$$\hat{\mathbf{D}}^* = \arg \min_{\mathcal{D} \in \mathbf{D}^+} \|\mathcal{W} - \mathcal{D}\|_F^2, \text{ and } \check{\mathbf{D}}^* = \arg \min_{\mathcal{D} \in \mathbf{D}^+} \|\mathcal{Z} - \mathcal{D}\|_F^2,$$

where $\mathbf{D}^+$ denotes the collection of all diagonal matrices with the nonnegative elements in $\mathbb{R}^{r \times r}$.

As known, measure functions provide helpful tools to evaluate and analyze well-conditioning of the square matrices. They often target the distribution of the matrix eigenvalues [25]. Among them, as a fundamental study to analyze the scaling and sizing of the quasi–Newton updates, Dennis and Wolkowicz [14] proposed the following measure function:

$$\psi(\mathbf{A}) = \frac{\text{tr}(\mathbf{A})}{r\sqrt[r]{\det(\mathbf{A})}}, \tag{2.5}$$

for an arbitrary positive definite matrix $\mathbf{A} \in \mathbb{R}^{r \times r}$. As a factor to evaluate well-conditioning, $\psi(\mathbf{A})$ considers all the eigenvalues of $\mathbf{A}$, rather than, as occurs in the spectral condition number, only taking the extreme eigenvalues of the matrix [30]. So,

by employing $\psi(.)$ instead of $\kappa(.)$ in (2.4), it is more likely possible to obtain NMF elements with well-distributed eigenvalues. However, the matrix function (2.5) would be accompanied by some kinds of complexity due to its denominator.

Mathematical inequalities have been widely and purposefully employed by the researchers to turn a dense or complicated formula into something manageable. For this aim, the first and foremost point in accordance with the norm of the literature is to rise the level of interpretability of the targeted formula or model. Here, for the sake of a well-planned simplicity that is a crucial issue in the high-dimensional data analysis, we organize assistance from the first part of the mean inequality chain that is related to the algebraic ties between the harmonic, geometric, arithmetic, and quadratic means. To proceed, firstly note that $\det(\mathbf{A}) = \prod_{i=1}^{r} \zeta_i$, in which $\{\zeta_i\}_{i=1}^{r}$ is the set of the eigenvalues of $\mathbf{A}$. Therefore, bearing the relation between the geometric and the harmonic means in mind, here in the sense of

$$\frac{r}{\sum_{i=1}^{r} \frac{1}{\zeta_i}} \leq \sqrt[r]{\prod_{i=1}^{r} \zeta_i},$$

and noting that the trace of a (square) matrix is equal to the sum of its eigenvalues, the following simple bound for $\psi(\mathbf{A})$ can be obtained:

$$\psi(\mathbf{A}) \leq \varphi(\mathbf{A}) = \frac{1}{r^2} \mathrm{tr}(\mathbf{A}) \mathrm{tr}(\mathbf{A}^{-1}).$$

This gives rise compelling motivations to employ $\varphi(.)$ instead of $\kappa(.)$ in (2.4), to possibly gain NMF elements with well-distributed eigenvalues. So, the modified model is given by

$$\begin{aligned} \breve{\mathfrak{F}}(\mathbf{W}, \mathbf{Z}) &= \frac{1}{2} \|\mathbf{A} - \mathbf{W}\mathbf{Z}\|_F^2 + \lambda_1 \varphi(\hat{\mathbf{D}}) + \lambda_2 \varphi(\check{\mathbf{D}}) \\ &= \frac{1}{2} \|\mathbf{A} - \mathbf{W}\mathbf{Z}\|_F^2 + \lambda_1 \frac{1}{r^2} \mathrm{tr}(\hat{\mathbf{D}}) \mathrm{tr}(\hat{\mathbf{D}}^{-1}) + \lambda_2 \frac{1}{r^2} \mathrm{tr}(\check{\mathbf{D}}) \mathrm{tr}(\check{\mathbf{D}}^{-1}) \\ &= \frac{1}{2} \|\mathbf{A} - \mathbf{W}\mathbf{Z}\|_F^2 + \frac{\lambda_1}{r^2} \left(\sum_{j=1}^{r} \sum_{i=1}^{m} \mathbf{W}_{ij}^2 \right) \left(\sum_{j=1}^{r} \frac{1}{\sum_{i=1}^{m} \mathbf{W}_{ij}^2} \right) \\ &\quad + \frac{\lambda_2}{r^2} \left(\sum_{i=1}^{r} \sum_{j=1}^{n} \mathbf{Z}_{ij}^2 \right) \left(\sum_{i=1}^{r} \frac{1}{\sum_{j=1}^{n} \mathbf{Z}_{ij}^2} \right). \end{aligned} \tag{2.6}$$

Inherited from the measure function (2.5), the penalty terms of the model (2.6) control the condition number by engaging in all the diagonal elements of the relevant matrices, not only considering the extreme ones. Also, emerging polynomial terms makes the model easier to handle with respect to determining the gradient of the cost function.

The major defect of the cost function of the model (2.6) is that it is not differentiable everywhere due to the extra penalty terms. Especially, if $\mathbf{W}$ has a zero column or $\mathbf{Z}$ has a zero row, then $\breve{\mathfrak{F}}$ in (2.6) is not well-defined. Moreover, small magnitudes of the columns of $\mathbf{W}$ or the rows of $\mathbf{Z}$ are computationally troublesome. Backed by these arguments and in favor of simplicity, our revised ANLS (RANLS) method is founded upon the following modified version of the model (2.6):

$$\tilde{\mathfrak{F}}(\mathbf{W},\mathbf{Z}) = \frac{1}{2}\|\mathbf{A}-\mathbf{W}\mathbf{Z}\|_F^2 + \frac{\lambda_1}{r^2}\left(\sum_{j=1}^{r}\sum_{i=1}^{m}\mathbf{W}_{ij}^2\right)\left(\sum_{j=1}^{r}\frac{1}{\gamma+\sum_{i=1}^{m}\mathbf{W}_{ij}^2}\right) + \frac{\lambda_2}{r^2}\left(\sum_{i=1}^{r}\sum_{j=1}^{n}\mathbf{Z}_{ij}^2\right)\left(\sum_{i=1}^{r}\frac{1}{\gamma+\sum_{j=1}^{n}\mathbf{Z}_{ij}^2}\right), \tag{2.7}$$

with some constant $\gamma > 0$. As seen, $\tilde{\mathfrak{F}}$ is well-defined and also, it is differentiable everywhere. Thus, the next revised versions of the least-squares models (2.2) and (2.3) should alternately be solved:

$$\mathbf{Z}^{k+1} = \arg\min_{\mathbf{Z}\geq 0}\tilde{\mathfrak{F}}(\mathbf{W}^k,\mathbf{Z}), \tag{2.8}$$

$$\mathbf{W}^{k+1} = \arg\min_{\mathbf{W}\geq 0}\tilde{\mathfrak{F}}(\mathbf{W},\mathbf{Z}^{k+1}), \tag{2.9}$$

for all $k \in \mathbb{Z}^+$.

3 Adaptive optimal choices for the Dai–Liao parameter based on the ellipsoid norm

Among the fundamental techniques for solving the unconstrained minimization problem $\min_{x\in\mathbb{R}^n} f(x)$, the CG methods are iteratively defined by

$$x_{k+1} = x_k + \alpha_k d_k,\ d_{k+1} = -g_{k+1} + \beta_k d_k,\ \forall k \in \mathbb{Z}^+, \tag{3.1}$$

starting by some $x_0 \in \mathbb{R}^n$ and $d_0 = -g_0$, in which $g_k = \nabla f(x_k)$ and $\beta_k \in \mathbb{R}$ is the CG parameter [3]. Also, the scalar $\alpha_k > 0$, called the step length, is customarily determined as the output of an approximate line search, popularly to meet the (strong) Wolfe conditions [28]. Here, we assume that the cost function f is smooth and its gradient is analytically available. Also, $\|.\|$ signifies the ℓ_2 (Euclidean) norm and our analysis undergoes with the Wolfe conditions for which $s_k^T y_k > 0$, where $s_k = x_{k+1} - x_k = \alpha_k d_k$.

In the initial years of the current century, the worthy study of Dai and Liao [13] brought considerable attention to the CG techniques in various guidelines [4]. Recently, Babaie–Kafaki [8] conducted an expository review on the DL method to provide a

better understanding of the capabilities of the method from several standpoints. For the DL method, β_k in its original form is set to

$$\beta_k^{\mathrm{DL}} = \frac{g_{k+1}^T y_k}{d_k^T y_k} - t\frac{g_{k+1}^T s_k}{d_k^T y_k}, \tag{3.2}$$

with $y_k = g_{k+1} - g_k$, where the scalar $t > 0$ is called the DL parameter. It is valuable to note that if

$$t \geq t_k^{(\eta)} := \eta\frac{\|y_k\|^2}{s_k^T y_k}, \text{ with the constant } \eta > \frac{1}{4}, \tag{3.3}$$

then the DL directions satisfy the sufficient descent condition which is an important ingredient of the convergence [5].

Among the analytical attempts to seek an appropriate formula for t as a classic open problem [4, 8], Babaie–Kafaki and Ghanbri [9] offered a least-squares model, i.e.

$$\min_{t>0} \|ts_k - y_k\|^2, \tag{3.4}$$

by pushing the DL direction to the direction of the three-term CG method proposed by Zhang, Zhou and Li (ZZL) [31]. As known, the ZZL directions satisfy a strong form of the sufficient descent condition. Moreover, they benefit the consecutive gradient differences vector y_k as an element of the search direction, besides the vectors g_{k+1} and d_k in the framework of a linear combination, rather than the DL directions that are just linear combination of g_{k+1} and d_k. As a result of their plan, Babaie–Kafaki and Ghanbri [9] obtained the following formula for t:

$$t := t_k^{\mathrm{ZZL}} = \frac{s_k^T y_k}{\|s_k\|^2}. \tag{3.5}$$

Here, we organize assistance from the ellipsoid vector norm to diversify the adaptive choices for the DL parameter. As an extended form of the ℓ_2 norm in the sense of

$$\|x\|_{\mathcal{M}} = \sqrt{x^T \mathcal{M} x},$$

where $\mathcal{M} \in \mathbb{R}^{n\times n}$ is a (symmetric) positive definite matrix, ellipsoid norm has been pivotally employed to analyze the convergence of the steepest descent and the quasi–Newton methods, and particularly, to devise the scaled trust region algorithms [28]. In our strategy, we plan to set several choices for $\mathcal{M}$ using the quasi–Newton updating formulas [28].

Quasi–Newton methods have been traditionally devised to tactfully estimate the (inverse) Hessian in order to determine the search direction in the iterative continuous optimization techniques. Mostly being positive definite, the given matrix approximations classically should satisfy the (standard) secant equation, i.e. $\mathbf{B}_{k+1}s_k = y_k$, where $\mathbf{B}_{k+1} \approx \nabla^2 f(x_{k+1})$ [3, 28]. The methods benefit enough flexibility to effectively address the large-scale models. For this aim, the memoryless versions of the

well-known BFGS and DFP quasi–Newton updating formulas can be applied [6]; that is,

$$\mathbf{H}_{k+1}^{\mathrm{MLBFGS}} = \mathbf{I} - \frac{y_k s_k^T + s_k y_k^T}{s_k^T y_k} + \left(1 + \frac{y_k^T y_k}{s_k^T y_k}\right) \frac{s_k s_k^T}{s_k^T y_k},$$

and

$$\mathbf{H}_{k+1}^{\mathrm{MLDFP}} = \mathbf{I} + \frac{s_k s_k^T}{s_k^T y_k} - \frac{y_k y_k^T}{y_k^T y_k},$$

both are positive definite approximations of $\nabla^2 f(x_{k+1})^{-1}$, for all $k \in \mathbb{Z}^+$. Here, MLBFGS and MLDFP are respectively shortened forms of the 'memoryless BFGS' and the 'memoryless DFP'.

It is a matter of tradition that reform is always needed in the available algorithms to answer the great need of diversity and inclusion. There are a lot of evidence in the methodology literature that such efforts evolved the algorithmic schemes over time. So, we should not neglect effects of these evolutionary plans on the hybrid CG algorithms as well. By this fact at the forefront, here we consider the ellipsoid extension of the least-squares model (3.4) as follows:

$$\min_{t>0} \|ts_k - y_k\|_{\mathcal{M}}^2,$$

which yields

$$t := t_k^{\mathcal{M}} = \frac{s_k^T \mathcal{M} y_k}{s_k^T \mathcal{M} s_k}.$$

So, t_k^{ZZL} given by (3.5) is the solution of (3.4) by setting $\mathcal{M}$ as the identity matrix. Also, if we let $\mathcal{M} = \mathbf{B}_{k+1}$ given by a quasi–Newton update for the Hessian, then, because of the standard secant equation we have

$$t := t_k^{\mathrm{DK}} = \frac{\|y_k\|^2}{s_k^T y_k}, \tag{3.6}$$

which is an effective formula already suggested by Dai and Kou (DK) [12]. This salient fact places great importance on the effectiveness of the given extended least-squares model. The setting $t = t_k^{\mathrm{DK}}$ in the DL method ensures (3.3) that squarely leads to the sufficient descent property. Moreover, if we let $\mathcal{M} = \mathbf{H}_{k+1}^{\mathrm{MLBFGS}}$ or $\mathcal{M} = \mathbf{H}_{k+1}^{\mathrm{MLDFP}}$, then we respectively obtain

$$t := t_k^{\mathrm{MLBFGS}} = \left(\frac{\|s_k\|^2}{s_k^T y_k} + \frac{\|y_k\|^2 \|s_k\|^2}{(s_k^T y_k)^2} - 1\right)^{-1}, \tag{3.7}$$

or

$$t := t_k^{\mathrm{MLDFP}} = \left(\frac{\|s_k\|^2}{s_k^T y_k} - \frac{(s_k^T y_k)^2}{\|y_k\|^2 \|s_k\|^2} + 1\right)^{-1}. \tag{3.8}$$

From the Cauchy–Schwarz inequality, it can be seen that $t_k^{\mathrm{MLBFGS}} > 0$ and $t_k^{\mathrm{MLDFP}} > 0$. To gain the sufficient descent property in light of (3.3), here we propose the following restricted versions of (3.7) and (3.8):

$$t_k^{\mathrm{MLBFGS}} \leftarrow \max\left\{t_k^{\mathrm{MLBFGS}}, t_k^{(\eta)}\right\}, \tag{3.9}$$

and

$$t_k^{\mathrm{MLDFP}} \leftarrow \max\left\{t_k^{\mathrm{MLDFP}}, t_k^{(\eta)}\right\}. \tag{3.10}$$

As a result, global convergence of the DL method with the given formulas for t can be proved following the analysis of [2, 13].

4 Computational experiments

We offer here some computational confirmation for the veracity of our theoretical analyses, starting with some numerical tests on the CUTEr library [18] with $n \geq 50$, comprising of 96 problems. All the tests were performed by MATLAB version 7.14.0.739 (R2012a), installed on the Centos 6.2 server Linux operation system, in a computer AMD FX–9800P RADEON R7 with 12 COMPUTE CORES 4C+8G 2.70 GHz of CPU and 8 GB of RAM. The effectuality of the parametric choices (3.6), (3.9), (3.10) and the Hager–Zhang (HZ) formula [19], i.e.

$$t := t_k^{\mathrm{HZ}} = 2\frac{\|y_k\|^2}{s_k^T y_k}, \tag{4.1}$$

is appraised for the DL+ method with

$$\beta_k^{\mathrm{DL+}} = \max\left\{\frac{g_{k+1}^T y_k}{d_k^T y_k}, 0\right\} - t\frac{g_{k+1}^T s_k}{d_k^T y_k}, \tag{4.2}$$

proposed for establishing convergence for general cost functions [13]. In our tests, DK+, DL–BFGS+, DL–DFP+ and HZ+, stand for the iterative method (3.1) with the CG parameter (4.2), in which t is respectively computed by (3.6), (3.7), (3.8) and (4.1). Since in rare iterations the DL+ method may fail to generate descent direction, restart (by the negative gradient vector) has been also employed as suggested in Dai and Liao [13].

For the algorithms, we used the approximate Wolfe conditions of Hager and Zhang [19] with the similar settings, and let the stopping criteria as $k > 10000$ or $\|g_k\| < 10^{-6}(1 + |f_k|)$. Also, we set $\eta = 0.26$ in (3.9) and (3.10), to enhance the possibility of employing (3.7) and (3.8). To visually assess the algorithmic results, we applied the Dolan–Moré performance profile [15], by comparisons based on the TNFGE and CPUT metrics, being acronyms for the 'total number of function and gradient evaluations' (as outlined in Hager and Zhang [19]) and the 'CPU time',

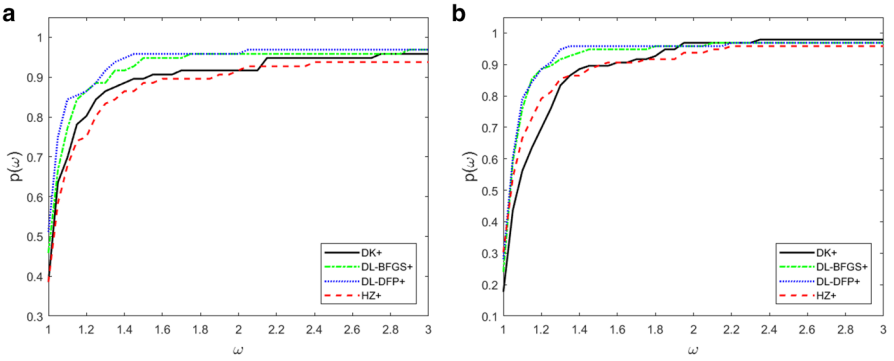

Fig. 1 Performance profile plots for DK+, DL–BFGS+, DL–DFP+ and HZ+ based on TNFGE (A) and CPUT (B)

respectively. Figure 1 represents the results, by which it can be seen that DL–BFGS+ and DL–DFP+ are generally preferable to DK+ and HZ+, especially with respect to TNFGE. Meanwhile, with respect to CPUT, at times DK+ and HZ+ are competitive with DL–BFGS+ and DL–DFP+. This observation is mainly related to the structure of the formulas (3.7) and (3.8) which is to some extent more complex rather than (3.6) and (4.1). Also, since DL–DFP+ is slightly preferable to DL–BFGS+, we can conclude that the setting (3.8) for the DL+ method is more effective than the setting (3.7).

To bring the validity of the given revised NMF model to light, in this part of our computational experiments we investigate the efficiency of DL–DFP+ for ANLS by solving the least-squares subproblem (2.2)–(2.3) of the minimization model (2.1), and for RANLS by solving the least-squares subproblem (2.8)–(2.9) of the minimization model (2.7). To handle the nonnegativity constraints in the subproblems, we followed the suggestion of Li et al. [22] and employed a proximal scheme in the sense of setting the negative entries of the iterative outputs equal to zero. For RALNS, we set $\lambda_1 = \lambda_2 = 1$ and $\gamma = 10^{-10}$ in (2.7), and for both ANLS and RANLS, we adopted the termination condition of Liu and Li [23] as well; which is

$$\|[\nabla_{\mathbf{Z}}\mathcal{F}(\mathbf{W}^k, \mathbf{Z}^k), \nabla_{\mathbf{W}}\mathcal{F}(\mathbf{W}^k, \mathbf{Z}^k)]\|_F \leq \nu\|[\nabla_{\mathbf{Z}}\mathcal{F}(\mathbf{W}^0, \mathbf{Z}^0), \nabla_{\mathbf{W}}\mathcal{F}(\mathbf{W}^0, \mathbf{Z}^0)]\|_F,$$

with $\mathcal{F} = \mathfrak{F}$ and $\mathcal{F} = \tilde{\mathfrak{F}}$, respectively, and $\nu = 10^{-2}$. By using the uniform distribution, the test matrices were generated randomly with various dimensions, together with the initial estimates of the NMF elements, as declared in Ahookhosh et al. [1]. Outputs have been outlined in Table 1, including the spectral condition number (Cond) and the relative error (RelErr), calculated by

$$\text{RelErr} = \frac{\|\mathbf{A} - \mathbf{WZ}\|_F}{\|\mathbf{A}\|_F}.$$

To recapitulate the results, we can observe that RANLS and ANLS are approximately competitive with respect to the accuracy. While, in the condition number

Table 1 The outputs of DL–DFP+ for NMF

Dimension (m, n, r)	Method	RelErr	Cond W	Cond Z
(50, 50, 4)	ALNS	6.90E-04	7.9068	5.3490
	RALNS	6.43E-04	3.4901	1.2639
(100, 50, 5)	ALNS	6.04E-04	5.6807	2.0834
	RALNS	5.85E-04	4.2506	1.2307
(100, 100, 5)	ALNS	7.11E-04	5.6165	2.2709
	RALNS	6.76E-04	4.2227	1.3666
(100, 250, 5)	ALNS	8.52E-04	5.1849	2.3579
	RALNS	7.31E-04	3.8486	1.3765
(200, 200, 4)	ALNS	7.18E-04	4.6986	2.5254
	RALNS	7.17E-04	3.4989	1.2648
(200, 200, 8)	ALNS	6.05E-04	22.5950	8.0738
	RALNS	7.61E-04	5.4005	1.4525
(200, 300, 6)	ALNS	6.93E-04	7.8248	3.1594
	RALNS	6.88E-04	4.5112	1.2894
(1000, 1000, 10)	ALNS	7.22E-04	216.8000	119.1700
	RALNS	2.50E-04	12.1330	4.6317
(2000, 3000, 12)	ALNS	2.93E-03	13.5770	51.4681
	RALNS	3.03E-03	3.6160	4.3847
(6000, 4000,10)	ALNS	1.11E-02	15.5430	12.9910
	RALNS	1.11E-02	13.0590	10.6700
(8000, 3000,15)	ALNS	1.82E-03	17.4270	7.4008
	RALNS	2.05E-03	15.8770	6.4922

viewpoint which is the main target of this study, RANLS is generally preferable to ANLS. Hence, capability of delivering well-conditioned NMF elements with satisfactory accuracy can therefore be considered a success by RANLS.

5 Conclusions

We have mainly concentrated on the modifying a classic optimization model of the nonnegative matrix factorization problem, frequently arising in a wide range of practical fields. Avoiding the possibility of ill-conditioning in the results of the decomposition motivated us to revise the model by embedding a measurement for condition numbers of the diagonalized types of the output matrices. What embedded as the well-conditioner (penalty) term has been extracted from the Dennis–Wolkowicz measure function [14]. Then, based on an ellipsoid norm-oriented least-squares model,

some optimal choices for the Dai–Liao parameter have been suggested. Driven by the great need for algorithmic tools with the low memory consumption of the machine, the ellipsoid norms have been centered on the memoryless BFGS and DFP formulas. The approach in terms of which the method's influential parameter has been computed is tending the Dai–Liao search direction to that of a well-functioning three-term conjugate gradient algorithm. Then, to examine the performance of the Dai–Liao method when it is equipped with the given formulas as the parametric settings, some computational tests were performed on the CUTEr functions. The findings were evaluated leveraged on the well-known Dolan–Moré benchmark. The results demonstrated the positive impact of our suggestions for the Dai–Liao parameter. Furthermore, the quality of the given revised nonnegative matrix factorization model has been assessed in several random cases. The results showed that the revised model can produce more well-conditioned factorization elements with reasonable relative errors. Thus, in practical terms, computational experiments have supported our theoretical assertions.

Acknowledgements The authors thank the anonymous reviewers for their worthy comments helped to improve the quality and organization of this work.

Author Contributions All authors whose names appear on the submission:
- made substantial contributions to the conception and design of the work, the acquisition, analysis, or interpretation of data;
- drafted the work or revised it critically for important intellectual content;
- approved the version to be published; and
- agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding Open access funding provided by Libera Universitá di Bolzano within the CRUI-CARE Agreement. No funds, grants, or other support was received.

Data Availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Ethical Approval The research does not involve human or animal participants. Also, the authors declare that they take on all the ethical responsibilities clarified by the journal.

Competing Interests The authors have no competing interests to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Ahookhosh, M., Hien, L.T.K., Gillis, N., Patrinos, P.: Multi-block Bregman proximal alternating linearized minimization and its application to orthogonal nonnegative matrix factorization. Comput. Optim. Appl. **79**(3), 681–715 (2021)
2. Aminifard, Z., Babaie-Kafaki, S.: An optimal parameter choice for the Dai–Liao family of conjugate gradient methods by avoiding a direction of the maximum magnification by the search direction matrix. 4OR. **17**, 317–330 (2019)
3. Andrei, N.: Modern Numerical Nonlinear Optimization. Switzerland, Springer, Cham (2006)
4. Andrei, N.: Open problems in conjugate gradient algorithms for unconstrained optimization. B. Malays. Math. Sci. So. **34**(2), 319–330 (2011)
5. Babaie-Kafaki, S.: On the sufficient descent condition of the Hager–Zhang conjugate gradient methods. 4OR. **12**(3), 285–292 (2014)
6. Babaie-Kafaki, S.: On optimality of the parameters of self-scaling memoryless quasi-Newton updating formulae. J. Optim. Theory Appl. **167**(1), 91–101 (2015)
7. Babaie-Kafaki, S.: A monotone preconditioned gradient method based on a banded tridiagonal inverse Hessian approximation. UPB Sci. Bull. Ser. A: Appl. Math. Phys. **80**(1), 55–62 (2018)
8. Babaie-Kafaki, S.: A survey on the Dai-Liao family of nonlinear conjugate gradient methods. RAIRO-Oper. Res. **57**, 43–58 (2023)
9. Babaie-Kafaki, S., Ghanbari, R.: Two optimal Dai-Liao conjugate gradient methods. Optimization **64**(11), 2277–2287 (2015)
10. Berry, M.W., Browne, M., Langville, A.N., Pauca, V.P., Plemmons, R.J.: Algorithms and applications for approximate nonnegative matrix factorization. Comput. Stat. Data Anal. **52**(1), 155–173 (2007)
11. Cho, Y.C., Choi, S.: Nonnegative features of spectro-temporal sounds for classification. Pattern Recognit. Lett. **26**(9), 1327–1336 (2005)
12. Dai, Y.H., Kou, C.X.: A nonlinear conjugate gradient algorithm with an optimal property and an improved Wolfe line search. SIAM J. Optim. **23**(1), 296–320 (2013)
13. Dai, Y.H., Liao, L.Z.: New conjugacy conditions and related nonlinear conjugate gradient methods. Appl. Math. Optim. **43**(1), 87–101 (2001)
14. Dennis, J.E., Wolkowicz, H.: Sizing and least-change secant methods. SIAM J. Numer. Anal. **30**(5), 1291–1314 (1993)
15. Dolan, E.D., Moré, J.J.: Benchmarking optimization software with performance profiles. Math. Program. **91**(2, Ser. A), 201–213 (2002)
16. Eldén, L.: Matrix Methods in Data Mining and Pattern Recognition. SIAM, Philadelphia (2007)
17. Gan, J., Liu, T., Li, L., Zhang, J.: Nonnegative matrix factorization: a survey. Comput. J. **64**(7), 1080–1092 (2021)
18. Gould, N.I.M., Orban, D., Toint, Ph.L.: CUTEr: a constrained and unconstrained testing environment, revisited. ACM Trans. Math. Software **29**(4), 373–394 (2003)
19. Hager, W.W., Zhang, H.: Algorithm 851: CG_Descent, a conjugate gradient method with guaranteed descent. ACM Trans. Math. Software **32**(1), 113–137 (2006)
20. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd Edition. Morgan Kaufmann (2012)
21. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8), 30–37 (2009)
22. Li, X., Zhang, W., Dong, X.: A class of modified FR conjugate gradient method and applications to nonnegative matrix factorization. Comput. Math. Appl. **73**, 270–276 (2017)
23. Liu, H., Li, X.: Modified subspace Barzilai-Borwein gradient method for non-negative matrix factorization. Comput. Optim. Appl. **55**(1), 173–196 (2013)
24. Pompili, F., Gillis, N., Absil, P.A., Glineur, F.: Two algorithms for orthogonal nonnegative matrix factorization with application to clustering. Neurocomputing **141**, 15–25 (2014)
25. Roozbeh, M., Babaie-Kafaki, S., Aminifard, Z.: Two penalized mixed-integer nonlinear programming approaches to tackle multicollinearity and outliers effects in linear regression models. J. Ind. Manag. Optim. **17**(6), 3475 (2021)
26. Roozbeh, M., Babaie-Kafaki, S., Aminifard, Z.: Improved high-dimensional regression models with matrix approximations applied to the comparative case studies with support vector machines. Optim. Methods Softw. **37**(5), 1912–1929 (2022)

27. Shahnaz, F., Berry, M.W., Pauca, V.P., Plemmons, R.J.: Document clustering using nonnegative matrix factorization. Inf. Process. Manag. **42**(2), 373–386 (2006)
28. Sun, W., Yuan, Y.X.: Optimization Theory and Methods: Nonlinear Programming. Springer, New York (2006)
29. Wang, Y., Jia, Y., Hu, C., Turk, M.: Nonnegative matrix factorization framework for face recognition. Int. J. Pattern Recognition Artif. Intell. **19**(04), 495–511 (2005)
30. Watkins, D.S.: Fundamentals of Matrix Computations. John Wiley and Sons, New York (2002)
31. Zhang, L., Zhou, W., Li, D.H.: Some descent three-term conjugate gradient methods and their global convergence. Optim. Methods Softw. **22**(4), 697–711 (2007)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.