ELSEVIER
Available online at www.sciencedirect.com
ScienceDirect
Linear Algebra and its Applications 422 (2007) 540–552
LINEAR ALGEBRA AND ITS APPLICATIONS
www.elsevier.com/locate/laa

Left vs right representations for solving weighted low-rank approximation problems

Ivan Markovsky [a], Sabine Van Huffel [b,*]

[a] *School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, United Kingdom*
[b] *Katholieke Universiteit Leuven, Departement Elektrotechniek, ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium*

Received 7 December 2005; accepted 13 November 2006
Available online 25 January 2007
Submitted by V. Mehrmann

Abstract

The weighted low-rank approximation problem in general has no analytical solution in terms of the singular value decomposition and is solved numerically using optimization methods. Four representations of the rank constraint that turn the abstract problem formulation into parameter optimization problems are presented. The parameter optimization problem is partially solved analytically, which results in an equivalent quadratically constrained problem. A commonly used re-parameterization avoids the quadratic constraint and makes the equivalent problem a nonlinear least squares problem, however, it might be necessary to change this re-parameterization during the iteration process. It is shown how the cost function can be computed efficiently in two special cases: row-wise and column-wise weighting.

Keywords: Weighted low-rank approximation; Total least squares; Parameter optimization

1. Introduction

We consider the following matrix approximation problem.

* Corresponding author. Tel.: +32 16 32 17 03; fax: +32 16 32 19 70.
E-mail addresses: ivan.markovsky@esat.kuleuven.be (I. Markovsky), sabine.vanhuffel@esat.kuleuven.be (S. Van Huffel).

0024-3795/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.laa.2006.11.012

Problem 1 (*Weighted low-rank approximation* (*WLRA*)). Given a matrix $D \in \mathbb{R}^{m\times n}$, a positive definite matrix $W \in \mathbb{R}^{mn\times mn}$, and an integer $r < \min(m,n)$, find an optimal W-weighted, rank-r approximation of D, defined as

$$\widehat{D}^* = \arg\min_{\widehat{D}} \operatorname{vec}^\top(D-\widehat{D})W\operatorname{vec}(D-\widehat{D}) \quad \text{subject to } \operatorname{rank}(\widehat{D}) \leqslant r. \tag{WLRA}$$

For $W = I$, the cost function of (WLRA) is $\|D-\widehat{D}\|_F^2$ and the problem has an analytical solution in terms of the singular value decomposition (SVD) of D.

Theorem 1 (Eckart–Young–Mirsky [2]). *Let $D = U\Sigma V^\top$ be the SVD of D and partition U, $\Sigma =: \operatorname{diag}(\sigma_1,\ldots,\sigma_q)$, $q := \min(m,n)$, and V as follows*:

$$U =: \begin{matrix} r & q-r & \\ \begin{bmatrix} U_1 & U_2 \end{bmatrix} & & m \end{matrix}, \quad \Sigma =: \begin{matrix} r & q-r & \\ \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} & & \begin{matrix} r \\ q-r \end{matrix} \end{matrix} \quad \text{and} \quad V =: \begin{matrix} r & q-r & \\ \begin{bmatrix} V_1 & V_2 \end{bmatrix} & & n \end{matrix},$$

Then the rank-r matrix

$$\widehat{D}^* = U_1\Sigma_1V_1^\top$$

is such that

$$\|D-\widehat{D}^*\|_F^2 = \min_{\operatorname{rank}(\widehat{D})\leqslant r} \|D-\widehat{D}\|_F^2 = \sigma_{r+1}^2 + \cdots + \sigma_q^2.$$

The solution $\widehat{D}^*$ is unique if and only if $\sigma_{r+1} \neq \sigma_r$.

Surprisingly, however, a similar analytical solution is not known for a general positive definite weight matrix W. Presently the most general WLRA problems that can be solved *exactly* using SVDs are those with a weight matrix of the form $W = W_\mathrm{r} \otimes W_\ell$, where $W_\mathrm{r} \in \mathbb{R}^{n\times n}$ and $W_\ell \in \mathbb{R}^{m\times m}$ are positive definite matrices and $\otimes$ is the Kronecker product. The solution procedure is based on the following theorem.

Theorem 2 (Two-sided WLRA). *Define the modified data matrix*

$$D_m := \sqrt{W_\ell}D\sqrt{W_\mathrm{r}},$$

and let $\widehat{D}_m^*$ be the optimal (*unweighted*) *low-rank approximation of D_m. Then*

$$\widehat{D}^* := \left(\sqrt{W_\ell}\right)^{-1}\widehat{D}_m^*\left(\sqrt{W_\mathrm{r}}\right)^{-1},$$

is a solution of the following two-sided WLRA problem

$$\min_{\widehat{D}} \|\sqrt{W_\ell}(D-\widehat{D})\sqrt{W_\mathrm{r}}\|_F^2 \quad \textit{subject to } \operatorname{rank}(\widehat{D}) \leqslant r.$$

A solution always exists. It is unique if and only if $\widehat{D}_m^*$ is unique.

The link between two-sided WLRA and (WLRA) is

$$\begin{aligned} \left\|\sqrt{W_\ell}(D-\widehat{D})\sqrt{W_\mathrm{r}}\right\|_F^2 &= \left\|\operatorname{vec}\left(\sqrt{W_\ell}(D-\widehat{D})\sqrt{W_\mathrm{r}}\right)\right\|^2 \\ &= \left\|\left(\sqrt{W_\mathrm{r}}\otimes\sqrt{W_\ell}\right)\operatorname{vec}(D-\widehat{D})\right\|^2 \\ &= \operatorname{vec}^\top(D-\widehat{D})\,(W_\mathrm{r}\otimes W_\ell)\,\operatorname{vec}(D-\widehat{D}). \end{aligned}$$

Here we used the identities

$$\mathrm{vec}(AXB) = (B^\top \otimes A)\mathrm{vec}(X) \quad \text{and} \quad (A_1 \otimes B_1)(A_2 \otimes B_2) = (A_1 A_2) \otimes (B_1 B_2).$$

Of course, $W \in \mathbb{R}^{mn \times mn}$, $m, n > 1$, being representable as $W_\mathrm{r} \otimes W_\ell$ with $W_\mathrm{r} \in \mathbb{R}^{n \times n}$ and $W_\ell \in \mathbb{R}^{m \times m}$ is a rather restrictive condition, i.e., representing a given positive definite weight matrix W in the form $W_\mathrm{r} \otimes W_\ell$, in general, requires approximation. It is shown in [10] that the optimal approximation (in Frobenius norm sense), i.e., the solution of the problem

$$\min_{W_\mathrm{r} \in \mathbb{R}^{n \times n}, W_\ell \in \mathbb{R}^{m \times m}} \|W - W_\mathrm{r} \otimes W_\ell\|_F \quad \text{subject to } W_\mathrm{r} \text{ and } W_\ell \text{ are positive definite,}$$

can be computed in analytical form in terms of the SVD of a matrix constructed from W. This result can be used for the computation of a suboptimal solution of the general WLRA problem as follows:

1. approximate the weight matrix W by the Kronecker product $W_\mathrm{r} \otimes W_\ell$ with some positive definite matrices $W_\mathrm{r} \in \mathbb{R}^{n \times n}$ and $W_\ell \in \mathbb{R}^{m \times m}$, using the result of [10], and
2. find an optimal WLRA of the data matrix D with the weight matrix $W_\mathrm{r} \otimes W_\ell$, using Theorem 2.

The computed WLRA with the weight matrix $W_\mathrm{r} \otimes W_\ell$ serves as a suboptimal solution for the desired WLRA with the weight matrix W. It is intuitively clear (and there is empirical evidence) that the better W is approximated, the closer the suboptimal solution to the optimal one is.

The suboptimal solution can be improved by numerical optimization methods. Since the problem is nonconvex there are no efficient optimization methods that are guaranteed to find a globally optimal solution. What is aimed at instead is a locally optimal solution nearby the given initial suboptimal approximation. Optimization methods for solving Problem 1 have been considered in the literature under different names:

- criss-cross multiple regression [3],
- Riemannian singular value decomposition [1],
- maximum likelihood principal component analysis [13],
- weighted low-rank approximation [4], and
- weighted total least squares [7,5].

Gabriel and Zamir [3] consider an element-wise weighted low-rank approximation problem, i.e., problem (WLRA) with diagonal weight matrix W, and propose an iterative solution method. Their method, however, does not necessarily converge to a minimum point, see the discussion in [3, Section 6, p. 491].

The Riemannian singular value decomposition framework of De Moor [1] includes the WLRA problem with rank specification $r = \min(m, n) - 1$ and a diagonal weight matrix W as a special case. In [1], an algorithm resembling the inverse power iteration algorithm is proposed. The method, however, has no proven convergence properties.

The maximum likelihood principal component analysis method of Wentzell et al. [13] is developed for applications in chemometrics, see also [9]. This method is an alternating least squares algorithm. It applies to the general WLRA problems and is globally convergent. The convergence rate, however, is linear and the method could be rather slow when the $r + 1$st and the rth singular values of the data matrix D are close to each other. In the unweighted case this situation corresponds to lack of uniqueness of the solution, cf., Theorem 1.

Manton et al. [4] treat the problem as an optimization over a Grassman manifold and propose steepest descent and Newton type algorithms. The least squares nature of the problem is not exploited in this work and the proposed algorithms are not globally convergent.

The method of [7,5] is a heuristic for solving the first order optimality condition of (WLRA). A solution of a nonlinear equation is sought instead of a minimum point of the original optimization problem. The method is locally convergent with super linear convergence rate. The method is not globally convergent and the region of convergence around a minimum point could be rather small in practice.

In [6, Chapter 3] it is shown that the above mentioned algorithms differ in

1. the way the rank constraint is represented and
2. the local optimization method applied to the resulting (from the particular representation used) parameter optimization problem.

Thus combining different rank representations with different local optimization methods, we obtain different solution methods for the WLRA problem.

An important remaining issue is the efficiency of the algorithm in special cases of the general WLRA approximation problem. Useful special cases are:

- *element-wise weighting*: $W = \operatorname{diag}(w_1, \dots, w_{mn})$, where $w_i > 0$, for $i = 1, \dots, mn$,
- *column-wise weighting*: $W = \operatorname{diag}(W_1, \dots, W_n)$, where $W_i \in \mathbb{R}^{m\times m}$, $W_i > 0$, for $i = 1, \dots, n$, and
- *row-wise weighting*: $T^\top W T = \operatorname{diag}(W_1, \dots, W_m)$, where $W_i \in \mathbb{R}^{n\times n}$, $W_i > 0$, for $i = 1, \dots, m$, and T is an $mn \times mn$ permutation matrix, such that $\operatorname{vec}(D) = T\operatorname{vec}(D^\top)$.

In this paper, we generalize the treatment of [6, Chapter 3], where left image and kernel representations and column-wise weighting, i.e., weight matrices $W = \operatorname{diag}(W_1, \dots, W_n)$, are considered. Our renewed interest in the problem was triggered from the difficulty that if $m > n$ the left image and kernel representations lead to larger number of parameters. It is advantageous to use in such cases right image and kernel representations. Note that the problem can *not* be resolved by transposing the data matrix because then W is in general no longer block diagonal and this structure is exploited in the computations.

We start from a full weight matrix W and then specialize the results to column and row-wise weighting. Section 2 presents four representations of the rank constraint in a parametric form. For these representations equivalent optimization problems are derived in Section 3 by eliminating the variable $\widehat{D}$. Section 4 presents unique parameterizations of the rank constraint, which simplifies the equivalent optimization problems. The resulting problems are nonlinear least squares problems and are solved by standard optimization methods, such as the Levenberg–Marquardt method.

2. Low-rank representations

In order to solve (either analytically or numerically) the WLRA problem, we represent the rank constraint $\operatorname{rank}(\widehat{D}) \leqslant r$ in a parametric form. Four possibilities are what are called image and kernel representations with their left and right versions. The left image representation

$$\operatorname{rank}(\widehat{D}) \leqslant r \Leftrightarrow \text{there is } P \in \mathbb{R}^{m\times r}, \quad P^\top P = I_r \text{ and } L \in \mathbb{R}^{r\times n}, \text{ such that } \widehat{D} = PL \tag{ℓIM}$$

imposes an upper bound on the dimension of the column space of $\widehat{D}$. This representation has as a parameter the matrix P. The right image representation

$$\operatorname{rank}(\widehat{D}) \leqslant r \Leftrightarrow \text{there is } P \in \mathbb{R}^{m\times r} \text{ and } L \in \mathbb{R}^{r\times n}, \quad LL^\top = I_r, \text{ such that } \widehat{D} = PL \tag{rIM}$$

has as a parameter the matrix L and imposes an upper bound on the dimension of the row space of $\widehat{D}$.

The left kernel representation

$$\operatorname{rank}(\widehat{D}) \leqslant r \Leftrightarrow \text{there is } R_\ell \in \mathbb{R}^{(m-r)\times m}, \quad R_\ell R_\ell^\top = I_{m-r}, \text{ such that } R_\ell \widehat{D} = 0 \tag{ℓKER}$$

has as a parameter the matrix R_ℓ and imposes a lower bound on the dimension of the left kernel of $\widehat{D}$. (The subscript "ℓ" in R_ℓ stands for "left".) The alternative (right) kernel representation

$$\operatorname{rank}(\widehat{D}) \leqslant r \Leftrightarrow \text{there is} R_\mathrm{r} \in \mathbb{R}^{n\times (n-r)}, \quad R_\mathrm{r}^\top R_\mathrm{r} = I_{n-r}, \text{ such that } \widehat{D} R_\mathrm{r} = 0 \tag{rKER}$$

has as a parameter the matrix R_r and imposes a lower bound on the dimension of the right kernel of $\widehat{D}$.

The four representations (ℓIM), (rIM), (ℓKER), and (rKER), however, lead to different number of parameters to be optimized over in solving (WLRA). Fewer optimization variables result in less computations per iteration as well as typically fewer iteration steps. Thus in the numerical solution of the WLRA problem it is important to use a representation that leads to as few as possible optimization variables.

Looking at the dimensions of the parameters P, L, R_ℓ, and R_r, we see that

- (ℓKER) has the fewest free parameters if $m < n$ and $r < m/2$,
- (ℓIM) has the fewest free parameters if $m < n$ and $r > m/2$,
- (rKER) has the fewest free parameters if $m > n$ and $r < n/2$, and
- (rIM) has the fewest free parameters if $m > n$ and $r > n/2$.

Due to the constraints $P^\top P = I$, $LL^\top = I$, $R_\ell R_\ell^\top = I$, and $R_\mathrm{r} R_\mathrm{r}^\top = I$, however, not all elements of the parameters are free. Equivalently the parameters P, L, R_ℓ, and R_r are not unique. In Section 4, we derive so-called left and right input/output parameterizations of which the parameters $X_\ell \in \mathbb{R}^{r\times(m-r)}$ and $X_\mathrm{r} \in \mathbb{R}^{r\times(n-r)}$ are unconstrained and unique.

3. Equivalent optimization problems

In what follows we use the notation

$$d := \operatorname{vec}(D) \quad \text{and} \quad \hat{d} := \operatorname{vec}(\widehat{D}).$$

Using the left image representation, (WLRA) yields the parameter optimization problem:

$$\min_{\substack{P \\ P^\top P = I}} \Bigg(\underbrace{\min_{\hat{d}, L} (d - \hat{d})^\top W (d - \hat{d}) \quad \text{subject to } \widehat{D} = PL}_{f_{\ell\text{IM}}(P)} \Bigg). \tag{WLRA$_P$}$$

Using the right image representation, (WLRA) yields the parameter optimization problem:

$$\min_{\substack{L \\ LL^\top = I}} \Bigg(\underbrace{\min_{\hat{d}, P} (d - \hat{d})^\top W (d - \hat{d}) \quad \text{subject to } \widehat{D} = PL}_{f_{\text{rIM}}(L)} \Bigg). \tag{WLRA$_L$}$$

Using the left kernel representation, (WLRA) yields the parameter optimization problem:

$$\min_{\substack{R_\ell \\ R_\ell R_\ell^\top = I}} \Bigg(\underbrace{\min_{\hat{d}} (d - \hat{d})^\top W (d - \hat{d}) \quad \text{subject to } R_\ell \widehat{D} = 0}_{f_{\ell\text{KER}}(R_\ell)} \Bigg). \tag{WLRA$_{R_\ell}$}$$

Using the right kernel representation, (WLRA) yields the parameter optimization problem:

$$\min_{\substack{R_\text{r} \\ R_\text{r}^\top R_\text{r} = I}} \Bigg(\underbrace{\min_{\hat{d}} (d - \hat{d})^\top W (d - \hat{d}) \quad \text{subject to } \widehat{D} R_\text{r} = 0}_{f_{\text{rKER}}(R_\text{r})} \Bigg). \tag{WLRA$_{R_\text{r}}$}$$

(WLRA$_P$), (WLRA$_L$), (WLRA$_{R_\ell}$), and (WLRA$_{R_\text{r}}$) are double minimization problems with outer minimization over the parameters of the rank representation and inner minimization over the data approximation. The inner minimization problems are convex quadratic optimization problems and can be solved analytically, yielding explicit expressions for the cost functions to be minimized over the parameters of the rank representation.

Define for a symmetric positive definite matrix $W \in \mathbb{R}^{mn \times mn}$ and a full column rank matrix $K \in \mathbb{R}^{mn \times \bullet}$

$$f_{\text{IM}}(K) := d^\top (W - WK(K^\top WK)^{-1} K^\top W) d \quad \text{and} \quad \hat{d}_{\text{IM}}(K) := K(K^\top WK)^{-1} K^\top W d. \tag{IM SOL}$$

Theorem 3 (Equivalent problems to (WLRA$_P$) and (WLRA$_L$)). *The optimization problem (WLRA$_P$) is equivalent to*

$$\min_P f_{\ell\text{IM}}(P) \quad \textit{subject to } P^\top P = I, \quad \textit{where } f_\ell\text{IM}(P) = f_{\text{IM}}(I_n \otimes P) \tag{WLRA$'_P$}$$

and the minimum is achieved at $\hat{d} = \hat{d}_{\text{IM}}(I_n \otimes P)$. *Similarly*, (WLRA$_P$) *is equivalent to*

$$\min_L f_{\text{rIM}}(L) \quad \textit{subject to } LL^\top = I, \quad \textit{where } f_{\text{rIM}}(L) = f_{\text{IM}}(L^\top \otimes I_m) \tag{WLRA$'_L$}$$

and the minimum is achieved at $\hat{d} = \hat{d}_{\text{IM}}(L^\top \otimes I_m)$.

Proof. For $f_{\ell\text{IM}}$, we have

$$\begin{aligned} f_{\ell\text{IM}} &= \min_{\hat{d}, L} (d - \hat{d})^\top W (d - \hat{d}) \quad \text{subject to } \hat{d} = \underbrace{(I_n \otimes P)}_{\mathbf{P}} \underbrace{\text{vec}(L)}_{l} \\ &= \min_l (d - \mathbf{P} l)^\top W (d - \mathbf{P} l). \end{aligned}$$

This is a standard convex quadratic optimization problem, of which the solution is (IM SOL) with $K = \mathbf{P} \in \mathbb{R}^{mn \times rn}$.

For f_{rIM}, we have

$$f_{\mathrm{rIM}} = \min_{\hat{d},P} (d - \hat{d})^\top W (d - \hat{d}) \quad \text{subject to } \hat{d} = \underbrace{(L^\top \otimes I_m)}_{\mathbf{L}} \underbrace{\mathrm{vec}(P)}_{p}$$
$$= \min_{p} (d - \mathbf{L}p)^\top W (d - \mathbf{L}p),$$

so that the solution is given by (IM SOL) with $K = \mathbf{L} \in \mathbb{R}^{mn \times mr}$. $\square$

Define for a symmetric positive definite matrix $W \in \mathbb{R}^{mn \times mn}$ and a full row rank matrix $K \in \mathbb{R}^{mn \times \bullet}$

$$f_{\mathrm{KER}}(K) := d^\top K^\top (K W^{-1} K^\top)^{-1} K d, \quad \text{and}$$
$$\hat{d}_{\mathrm{KER}}(K) := \left(I - W^{-1} K^\top (K W^{-1} K^\top)^{-1} K\right) d. \qquad \text{(KER SOL)}$$

Theorem 4 (Equivalent problems to (WLRA$_{R_\ell}$) and (WLRA$_{R_\mathrm{r}}$)). *The optimization problem (WLRA$_{R_\ell}$) is equivalent to*

$$\min_{R_\ell} f_{\ell\mathrm{KER}}(R_\ell) \quad \textit{subject to } R_\ell R_\ell^\top = I, \quad \textit{where } f_{\ell\mathrm{KER}(R_\ell) = f_{\mathrm{KER}}}(I_n \otimes R_\ell) \qquad (\mathrm{WLRA}'_{R\ell})$$

and the minimum is achieved at $\hat{d} = \hat{d}_{\mathrm{KER}}(I_n \otimes R_\ell)$. *Similarly, (WLRA$_{R_\mathrm{r}}$) is equivalent to*

$$\min_{R_\mathrm{r}} f_{\mathrm{rKER}}(R_\mathrm{r}) \quad \textit{subject to } R_\mathrm{r}^\top R_\mathrm{r} = I, \quad \textit{where } f_{\mathrm{rKER}}(R_\mathrm{r}) = f_{\mathrm{KER}}(R_\mathrm{r}^\top \otimes I_m) \qquad (\mathrm{WLRA}'_{R\mathrm{r}})$$

and the minimum is achieved at $\hat{d} = \hat{d}_{\mathrm{KER}}(R_\mathrm{r}^\top \otimes I_m)$.

Proof. Let $\Delta D := D - \widehat{D}$ and $\Delta d := \mathrm{vec}(\Delta D)$. For $f_{\ell\mathrm{KER}}$, we have

$$f_{\ell\mathrm{KER}} = \min_{\Delta d} \Delta d^\top W \Delta d \quad \text{subject to } R_\ell (D - \Delta D) = 0$$
$$= \min_{\Delta d} \Delta d^\top W \Delta d \quad \text{subject to } (I_n \otimes R_\ell) \Delta d = \mathrm{vec}(R_\ell D),$$

which is a standard weighted least norm problem. The solution is given by (KER SOL) with $K = I_n \otimes R_\ell$. For f_{rKER}, we have

$$f_{\mathrm{rKER}} = \min_{\Delta d} \Delta d^\top W \Delta d \quad \text{subject to } (D - \Delta D) R_\mathrm{r} = 0$$
$$= \min_{\Delta d} \Delta d^\top W \Delta d \quad \text{subject to } (R_\mathrm{r} \otimes I_m) \Delta d = \mathrm{vec}(D R_\mathrm{r}),$$

of which the solution is given by (KER SOL) with $K = R_\mathrm{r} \otimes I_m$. $\square$

4. Free parameters and input/output parameterizations

The outer minimization problems (WLRA$'_L$), (WLRA$'_P$), (WLRA$'_{R_\ell}$), and (WLRA$'_{R_\mathrm{r}}$) are difficult constrained nonconvex optimization problems. In order to simplify them, we first eliminate the quadratic constraints $P^\top P = I$, $LL^\top = I$, $R_\ell R_\ell^\top = I$, and $R_\mathrm{r}^\top R_\mathrm{r} = I$ by re-parameterizing the problems. The parameters P, L, R_ℓ, and R_r are not unique and the nonuniqueness is up to a choice of basis for col span(P), row span(L), ker(R_ℓ), and leftker(R_r), respectively.

Theorem 5 (Equivalent parameters). *For any nonsingular matrices U, V, Q_ℓ, and Q_r of appropriate size*,

- $f_{\ell\mathrm{IM}}(P) = f_{\ell\mathrm{IM}}(PU)$,
- $f_{\mathrm{rIM}}(L) = f_{\mathrm{rIM}}(VL)$,
- $f_{\ell\mathrm{KER}}(R_\ell) = f_{\ell\mathrm{KER}}(Q_\ell R_\ell)$, *and*
- $f_{\mathrm{rKER}}(R_\mathrm{r}) = f_{\mathrm{rKER}}(R_\mathrm{r} Q_\mathrm{r})$.

In the above sense the parameters PU, VL, $Q_\ell R_\ell$, and $R_\mathrm{r} Q_\mathrm{r}$ are equivalent to P, L, R_ℓ, and R_r, respectively.

Proof. By the properties of the Kronecker product, we have

- $I_n \otimes (PU) = (I_n \otimes P)(I_n \otimes U)$,
- $(VL)^\top \otimes I_m = (L^\top \otimes I_m)(V^\top \otimes I_m)$,
- $I_n \otimes (Q_\ell R_\ell) = (I_n \otimes Q_\ell)(I_n \otimes R_\ell)$,
- $(R_\mathrm{r} Q_\mathrm{r})^\top \otimes I_m = (Q_\mathrm{r}^\top \otimes I_m)(R_\mathrm{r}^\top \otimes I_m)$.

The matrices $I_n \otimes U$, $V^\top \otimes I_m$, $I_n \otimes R_\ell$, $R_\mathrm{r}^\top \otimes I_m$ are nonsingular if and only if, respectively, the matrices U, V, Q_ℓ, Q_r are nonsingular.

Consider the solution (IM SOL) for the problems with image representations and let T be a nonsingular matrix of dimension coldim(K). We will show that $f_{\mathrm{IM}}(K) = f_{\mathrm{IM}}(KT)$. Indeed,

$$\begin{aligned} d_{\mathrm{IM}}(KT) &:= KT(T^\top K^\top W K T)^{-1} T^\top K^\top W d \\ &= KTT^{-1}(K^\top W K)^{-1}(T^\top)^{-1} T^\top K^\top W d = d_{\mathrm{IM}}(K). \qquad \square \end{aligned}$$

Consider then the solution (KER SOL) for the problems with kernel representations and let T be a nonsingular matrix of dimension row dim(K). We will show that $f_{\mathrm{KER}}(K) = f_{\mathrm{KER}}(TK)$. Indeed,

$$\begin{aligned} d_{\mathrm{KER}}(TK) &:= (I - W^{-1} K^\top T^\top (TKW^{-1}K^\top T^\top)^{-1} TK) d \\ &= (I - W^{-1} K^\top T^\top (T^\top)^{-1} (KW^{-1}K^\top) T^{-1} TK) d = d_{\mathrm{KER}}(K). \qquad \square \end{aligned}$$

Next we show special equivalent parameters that suggest a re-parameterization eliminating the nonuniqueness. Let $\Pi_\ell \in \mathbb{R}^{m\times m}$ and $\Pi_\mathrm{r} \in \mathbb{R}^{n\times n}$ be permutation matrices and define the partitionings

$$\Pi_\ell P =: \overset{r}{\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}} \begin{matrix} r \\ m-r \end{matrix}, \quad L\Pi_\mathrm{r} =: \begin{bmatrix} \overset{r}{L_1} & \overset{n-r}{L_2} \end{bmatrix} r, \quad R_\ell \Pi_\ell^{-1} =: \begin{bmatrix} \overset{r}{R_{\ell,1}} & \overset{m-r}{R_{\ell,2}} \end{bmatrix} m-r,$$

$$\Pi_\mathrm{r}^{-1} R_\mathrm{r} =: \overset{n-r}{\begin{bmatrix} R_{\mathrm{r},1} \\ R_{\mathrm{r},2} \end{bmatrix}} \begin{matrix} r \\ n-r \end{matrix}. \qquad \text{(I/O PRT)}$$

Since P and R_r are full column rank and L and R_ℓ are full row rank, then there are permutations Π_ℓ and Π_r, such that the square blocks P_1, L_1, $R_{\ell,2}$, and $R_{\mathrm{r},2}$ are nonsingular. In what follows, assume that Π_ℓ and Π_r are such permutations. Then equivalent parameters to $\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$, $\begin{bmatrix} L_1 & L_2 \end{bmatrix}$, $\begin{bmatrix} R_{\ell,1} & R_{\ell,2} \end{bmatrix}$, and $\begin{bmatrix} R_{\mathrm{r},1} \\ R_{\mathrm{r},2} \end{bmatrix}$ are

$$\widetilde{P} := \begin{bmatrix} I_r \\ P_2 P_1^{-1} \end{bmatrix}, \quad \widetilde{L} := \begin{bmatrix} I_r & L_1^{-1} L_2 \end{bmatrix}, \quad \widetilde{R}_\ell := \begin{bmatrix} -R_{\ell,2}^{-1} R_{\ell,1} & -I_{m-r} \end{bmatrix},$$
$$\widetilde{R}_\mathrm{r} := \begin{bmatrix} -R_{\mathrm{r},1} R_{\mathrm{r},2}^{-1} \\ -I_{n-r} \end{bmatrix}.$$

Theorem 6 (Link between image and kernel parameters). *If col span $(P) = ker(R_\ell)$ and at least one of the blocks P_1 or $R_{\ell,2}$, defined in (I/O PRT), is nonsingular, then the other block is also nonsingular and*

$$P_2 P_1^{-1} = -R_{\ell,2}^{-1} R_{\ell,1} =: X_\ell \in \mathbb{R}^{(m-r)\times r}.$$

Similarly, if row span(L) = left ker(R_r) *and at least one of the blocks L_1 or $R_{\mathrm{r},2}$, defined in (I/O PRT), is nonsingular, then the other block is also nonsingular and*

$$L_1^{-1} L_2 = -R_{\mathrm{r},1} R_{\mathrm{r},2}^{-1} =: X_\mathrm{r}. \in \mathbb{R}^{r\times(n-r)}.$$

Proof. For the first statement we have

$$\text{col span}(P) = \ker(R_\ell) \Rightarrow \widetilde{R}_\ell \widetilde{P} = 0 \Rightarrow \begin{bmatrix} -R_{\ell,2}^{-1} R_{\ell,1} & -I_{m-r} \end{bmatrix} \begin{bmatrix} I_r \\ P_2 P_1^{-1} \end{bmatrix}$$
$$= 0 \Rightarrow P_2 P_1^{-1} = -R_{\ell,2}^{-1} R_{\ell,1}$$

and for the second one

$$\text{row span}(L) = \text{leftker}(R_\mathrm{r}) \Rightarrow \widetilde{L}\widetilde{R}_\mathrm{r} = 0 \Rightarrow \begin{bmatrix} I_r & L_1^{-1} L_2 \end{bmatrix} \begin{bmatrix} -R_{\mathrm{r},1} R_{\mathrm{r},2}^{-1} \\ -I_{n-r} \end{bmatrix}$$
$$\Rightarrow L_1^{-1} L_2 = -R_{\mathrm{r},1} R_{\mathrm{r},2}^{-1}. \qquad \square$$

Define the two partitionings of the matrix $\widehat{D}$

$$\Pi_\ell \widehat{D} =: \begin{bmatrix} \widehat{A}_\ell \\ \widehat{B}_\ell \end{bmatrix} \begin{matrix} r \\ m-r \end{matrix} \quad \text{and} \quad \widehat{D}\Pi_\mathrm{r} =: \begin{matrix} r & n-r \\ \end{matrix} \begin{bmatrix} \widehat{A}_\mathrm{r} & \widehat{B}_\mathrm{r} \end{bmatrix}.$$

The representations

$$\text{there is } X_\ell \in \mathbb{R}^{(m-r)\times r}, \text{ such that } X_\ell \widehat{A}_\ell = \widehat{B}_\ell \Rightarrow \operatorname{rank}(\widehat{D}) \leqslant r \tag{lI/O}$$

and

$$\text{there is } X_\mathrm{r} \in \mathbb{R}^{r\times(n-r)}, \text{ such that } \widehat{A}_\mathrm{r} X_\mathrm{r} = \widehat{B}_\mathrm{r} \Rightarrow \operatorname{rank}(\widehat{D}) \leqslant r \tag{rI/O}$$

are called input/output ones. Their parameters X_ℓ and X_r are unconstrained. Using the left input/output parameterization (lI/O), the WLRA problem becomes

$$\min_{\widehat{A}_\ell, \widehat{B}_\ell, X_\ell} \operatorname{vec}^\top \left(D - \Pi_\ell^{-1} \begin{bmatrix} \widehat{A}_\ell \\ \widehat{B}_\ell \end{bmatrix} \right) W \operatorname{vec} \left(D - \Pi_\ell^{-1} \begin{bmatrix} \widehat{A}_\ell \\ \widehat{B}_\ell \end{bmatrix} \right) \quad \text{subject to } X_\ell \widehat{A}_\ell = \widehat{B}_\ell \tag{lWTLS}$$

and using the right input/output parameterization (rI/O), the WLRA problem becomes

$$\min_{\widehat{A}_\mathrm{r},\widehat{B}_\mathrm{r},X_\mathrm{r}} \operatorname{vec}^\top\left(D-\begin{bmatrix}\widehat{A}_\mathrm{r} & \widehat{B}_\mathrm{r}\end{bmatrix}\Pi_\mathrm{r}^{-1}\right)W\operatorname{vec}\left(D-\begin{bmatrix}\widehat{A}_\mathrm{r} & \widehat{B}_\mathrm{r}\end{bmatrix}\Pi_\mathrm{r}^{-1}\right)$$
$$\text{subject to } \widehat{A}_\mathrm{r}X_\mathrm{r}=\widehat{B}_\mathrm{r}. \qquad \text{(rWTLS)}$$

For fixed permutation matrices, problems (lWTLS) and (rWTLS) are known in the literature as weighted total least squares problems. Under the assumption that the permutations Π_ℓ and Π_r are properly chosen, (lWTLS) and (rWTLS) are equivalent to (WLRA).

Note 1. If the permutations Π_ℓ and Π_r are not properly chosen, the weighted total least squares problems might have no solution or be close to having no solution (which leads to ill conditioned computational problems). Such cases are called in the literature nongeneric total least squares problems and have been analyzed in [11,12,8]. In particular, the formulation of [8] defines a so-called core problem, corresponding to a given total least squares problem, that always has a unique solution. The solution of the core problem coincides with the classical total least squares solution whenever the latter exists and leads to the nongeneric total least squares solution, defined in [11,12], when the classical total least squares solution fails to exist. The problem of detecting when a given problem is nongeneric, however, involves checking singularity of a submatrix and is therefore not straightforward numerically. In addition, the core problem is currently formulated only for single right-hand-side problems (corresponding to low-rank approximation problems with rank reduction by one). Statistical methods for detecting nongeneric cases and generalization of the core problem formulation for the multiple-right-hand side total least squares problems are topics of current research.

We solve the unconstrained outer minimization problems

$$\begin{aligned}&\min\nolimits_{X_\ell} f_{\ell\mathrm{IM}}\left(\Pi_\ell^{-1}\begin{bmatrix}I\\ X_\ell\end{bmatrix}\right), \quad \min\nolimits_{X_\mathrm{r}} f_{\mathrm{rIM}}\left(\begin{bmatrix}I & X_\mathrm{r}\end{bmatrix}\Pi_\mathrm{r}^{-1}\right),\\ &\min\nolimits_{X_\ell} f_{\ell\mathrm{KER}}\left(\begin{bmatrix}X_\ell & -I\end{bmatrix}\Pi_\mathrm{r}\right), \quad \text{and} \quad \min\nolimits_{X_\mathrm{r}} f_{\mathrm{rKER}}\left(\Pi_\ell\begin{bmatrix}X_\mathrm{r}\\ -I\end{bmatrix}\right)\end{aligned} \qquad \text{(WLRA}''\text{)}$$

using local optimization methods for nonlinear least squares problems, e.g., the Levenberg–Marquardt method. Applying these methods, of prime importance is the number of optimization parameters and the efficiency of the cost function (and its derivatives) evaluation. From the point of view of minimal number of optimization variables the left (right) input/output parameterization should be chosen when $m < n$ ($m > n$), irrespective of the rank reduction r.

Next we verify on simulation examples that indeed the optimization problem using the "correct" (left or right) input/output representation is easier to solve in the sense that it requires fewer iterations steps under the same conditions (initial approximation and convergence tolerance).

Simulation example

The data matrix D is a sum of a random rank-r matrix and a random full rank matrix (errors-in-variables setup). The weight matrix W is a random positive definite matrix. In MATLAB D and W are generated as follows:

```
d0 = rand(m, r) * rand(r, n);
d0 = d0/norm(d0,'fro');              %normalization
```

Table 1
Number of optimization variables and average iterations for the optimization algorithm solving (WLRA″) by left and right kernel representations

	m n	2 10	3 9	4 8	5 7	6 6	7 5	8 4	9 3	10 2
$r=1$	Left	1/4	2/6	3/7	4/10	5/5	6/18	7/9	8/10	9/11
	Right	9/11	8/13	7/7	6/10	5/5	4/11	3/7	2/6	1/4
$r=2$	Left	–	2/5	4/27	6/21	8/11	10/44	12/27	14/28	–
	Right	–	14/12	12/34	10/69	8/11	6/39	4/24	2/6	–
$r=3$	Left	–	–	3/11	6/26	9/11	12/46	15/53	–	–
	Right	–	–	15/30	12/33	9/14	6/44	3/10	–	–
$r=4$	Left	–	–	–	4/9	8/14	16/12	–	–	–
	Right	–	–	–	16/12	8/13	4/8	–	–	–

```
 w = rand(m * n, 2 * m * n); w = w * w';
dt = reshape(sqrtm(w)\randn(m * n, 1), m, n);
dt = dt/norm(dt,'fro');             %normalization
 d = d0 + 0.001 * dt;
```

The matrix dimension m is varied from 2 to 10 and $n := 12 - m$. The desired rank r is varied from 1 to 4. For each combination of m and r, $N = 50$ independently generated data matrices are generated and for each one of them the problems (WLRA″) with left and right input/output representations are solved.

We use the Levenberg–Marquardt optimization algorithm from the Optimization toolbox of MATLAB (`lsqnonlin`) with cost function evaluations only. In both cases (left and right input/output representation), the algorithm is run from equivalent initial approximations obtained using the singular value decompositions (unweighted low-rank approximation) and the stopping criteria are set to the same tolerances. In this particular experiment, in all runs the same solution was found by using the two representations. As expected the average number of iterations, however, depends on the number of optimization parameters (elements of X_ℓ and X_r): the fewer the optimization variables, the fewer the average number of iterations for convergence. Numerical results are shown in Table 1 in the format "# of optimization variables/# of iterations".

5. Special cases of column and row-weighting

Up to now we have considered the WLRA problem with a full weight matrix W. Of interest, however, are the two special cases of element-wise weighting, i.e., W diagonal, column-wise weighting, i.e., W block diagonal with blocks of dimension $m \times m$, and row-wise weighting, i.e., $T^\top W T$ block diagonal with blocks of dimension $n \times n$, where T is the $mn \times mn$ permutation matrix, such that $\mathrm{vec}(D) = T\mathrm{vec}(D^\top)$. The special structure of W can be exploited for efficient cost function and first derivative evaluation.

Consider the column-wise weighting case, i.e., W is block-diagonal with blocks of dimension $m \times m$. It turns out that the cost function formulas using the left image and left kernel representations are easy to rewrite taking into account the special structure of W.

$$\hat{d}_{\ell\text{IM}} := (I \otimes P)((I \otimes P)^\top W (I \otimes P))^{-1}(I \otimes P)^\top W d$$
$$\Leftrightarrow \hat{d}_{\ell\text{IM},i} = P(P^\top W_i P)^{-1} P^\top W_i d_i,$$

where d_i is the ith column of D and $\hat{d}_i$ is the ith column of $\widehat{D}$. Similar decoupling into independent terms occurs for the left kernel representation:

$$\hat{d}_{\ell\text{KER}} = \Big(I_{mn} - W^{-1}(I \otimes R_\ell)^\top \Big((I \otimes R_\ell) W (I \otimes R_\ell)^\top\Big)^{-1} (I \otimes R_\ell)\Big) d$$
$$\Leftrightarrow \hat{d}_{\ell\text{KER},i} = \Big(I_m - W_i^{-1} R_\ell^\top (R_\ell W_i R_\ell^\top)^{-1} R_\ell\Big) d_i.$$

In the case of right image and right kernel representations the computational savings are achieved as follows. For the right image representation

$$\hat{d}_{\text{rIM}} := (L^\top \otimes I_m)\Big((L^\top \otimes I_m)^\top W (L^\top \otimes I_m)\Big)^{-1} (L^\top \otimes I_m)^\top W d, \tag{*}$$

denoting the ith column of L by l_i, we have

$$(L^\top \otimes I_m)^\top W (L^\top \otimes I_m) = \begin{bmatrix} l_1 \otimes I_m & \cdots & l_n \otimes I_m \end{bmatrix} \times \begin{bmatrix} W_1 & & \\ & \ddots & \\ & & W_n \end{bmatrix} \begin{bmatrix} l_1^\top \otimes I_m \\ \vdots \\ l_n^\top \otimes I_m \end{bmatrix} = \sum_{i=1}^{n} l_i l_i^\top \otimes W_i.$$

Let

$$p := \left(\sum_{i=1}^{n} l_i l_i^\top \otimes W_i\right)^{-1} \begin{bmatrix} l_1 \otimes W_1 & \cdots & l_n \otimes W_n \end{bmatrix} d \tag{**}$$

and $P := \begin{bmatrix} p_1 & \cdots & p_r \end{bmatrix}$, where $p^\top =: \begin{bmatrix} p_1^\top & \cdots & p_r^\top \end{bmatrix}$ and $p_i \in \mathbb{R}^m$. Note that $p = \text{vec}(P)$. Finally,

$$\hat{d}_{\text{rIM}} = (L^\top \otimes I_m) p \Rightarrow \widehat{D}_{\text{rIM}} = PL.$$

Computing p using (**) and setting $\widehat{D}_{\text{rIM}} = PL$ is a more efficient alternative to the direct computation of $\hat{d}_{\text{rIM}}$ using (*).

6. Conclusions

We have presented four representations of the rank constraint that turn the abstract weighted low-rank problem formulation in concrete parameter optimization problems. The optimization problem was solved analytically for the approximating matrix $\widehat{D}$, which resulted in an equivalent optimization problem over the parameters of the rank representation only. We showed a unique parameterization that avoids the quadratic constraint and further transforms the optimization problem into a classical nonlinear least squares problem. Finally we showed how the cost function can be computed efficiently in two special cases: row-wise and column-wise weighting.

Acknowledgments

Our research is supported by Research Council KUL: GOA-AMBioRICS, GOA-Mefisto 666, Center of Excellence EF/05/006 "Optimization in engineering", several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects, G.0360.05 (EEG signal processing), G.0321.06 (numerical tensor techniques), research communities (ICCoS, ANMMM); IWT: PhD Grants; Belgian Federal Science Policy Office IUAP P5/22 ('Dynamical Systems and Control: Computation, Identification and Modelling'); EU: BIOPATTERN, ETUMOUR; HEALTHagents.

References

[1] B. De Moor, Structured total least squares and L_2 approximation problems, Linear Algebra Appl. 188–189 (1993) 163–207.
[2] G. Eckart, G. Young, The approximation of one matrix by another of lower rank, Psychometrika 1 (1936) 211–218.
[3] K. Gabriel, S. Zamir, Lower rank approximation of matrices by least squares with any choice of weights, Technometrics 21 (1979) 489–498.
[4] J. Manton, R. Mahony, Y. Hua, The geometry of weighted low-rank approximations, IEEE Trans. Signal Process. 51 (2) (2003) 500–514.
[5] I. Markovsky, M.-L. Rastello, A. Premoli, A. Kukush, S. Van Huffel, The element-wise weighted total least squares problem, Comput. Statist. Data Anal. 50 (1) (2005) 181–209.
[6] I. Markovsky, J.C. Willems, S. Van Huffel, B. De Moor, Exact and Approximate Modeling of Linear Systems: A Behavioral Approach, Monographs on Mathematical Modeling and Computation, vol. 11, SIAM, 2006.
[7] A. Premoli, M.-L. Rastello, The parametric quadratic form method for solving TLS problems with elementwise weighting, in: S. Van Huffel, P. Lemmerling (Eds.), Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications, Kluwer, 2002, pp. 67–76.
[8] C. Paige, Z. Strakos, Core problems in linear algebraic systems, SIAM J. Matrix Anal. Appl. 27 (2005) 861–875.
[9] M. Schuermans, I. Markovsky, P. Wentzell, S. Van Huffel, On the equivalence between total least squares and maximum likelihood PCA, Anal. Chim. Acta 544 (2005) 254–267.
[10] C. Van Loan, N. Pitsianis, Approximation with Kronecker products, in: M. Moonen, G. Golub (Eds.), Linear Algebra for Large Scale and Real Time Applications, Kluwer Publications, 1993, pp. 293–314.
[11] S. Van Huffel, J. Vandewalle, Analysis and solution of the nongeneric total least squares problem, SIAM J. Matrix Anal. Appl. 9 (1988) 360–372.
[12] S. Van Huffel, J. Vandewalle, The Total Least Squares Problem: Computational Aspects and Analysis, SIAM, Philadelphia, 1991.
[13] P. Wentzell, D. Andrews, D. Hamilton, K. Faber, B. Kowalski, Maximum likelihood principle component analysis, J. Chemometrics 11 (1997) 339–366.