# Approximate low-rank factorization with structured factors[☆]

Ivan Markovsky[*], Mahesan Niranjan

*School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK*

## Abstract

An approximate rank revealing factorization problem with structure constraints on the normalized factors is considered. Examples of structure, motivated by an application in microarray data analysis, are sparsity, nonnegativity, periodicity, and smoothness. In general, the approximate rank revealing factorization problem is nonconvex. An alternating projections algorithm is developed, which is globally convergent to a locally optimal solution. Although the algorithm is developed for a specific application in microarray data analysis, the approach is applicable to other types of structure.

*Key words:* rank revealing factorization; numerical rank; low-rank approximation; maximum likelihood PCA; total least squares; errors-in-variables; microarray data.

## 1. Introduction

*Rank estimation*

Consider an $m \times n$ real matrix $X_0$ with rank $r_0 < \min(m,n)$. A factorization $X_0 = C_0 P_0$, where $C_0$ is $m \times r_0$ and $P_0$ is $r_0 \times n$ is called *rank revealing*. Suppose that instead of $X_0$ a matrix $X := X_0 + E$ is observed, where $E$ is a perturbation, e.g., $E$ can represent rounding errors in a finite precision arithmetic or measurement errors in data acquisition. The rank of the perturbed matrix $X$ may not be equal to $r_0$. If $E$ is random, generically, $X$ is full rank, so that from a practical point of view, a nonzero perturbation $E$ makes the matrix $X$ full rank. If, however, $E$ is "small", in the sense that its Frobenius norm $\|E\|_\text{F} := \sqrt{\sum_{i=1}^m \sum_{j=1}^n e_{ij}^2}$ is less than a constant $\varepsilon$ (defining the perturbation size), then $X$ will be "close" to a rank-$r_0$ matrix in the sense that the distance of $X$ to the manifold of rank-$r_0$ matrices

$$d(X, r_0) := \min_{\widehat{X}} \quad \|X - \widehat{X}\|_\text{F} \quad \text{subject to} \quad \text{rank}(\widehat{X}) = r_0 \tag{1}$$

[☆] MATLAB files reproducing the numerical results and figures presented in the paper are available from: `http://users.ecs.soton.ac.uk/im/factorize.tar`

[*] Corresponding author

*Email addresses:* `im@ecs.soton.ac.uk` (Ivan Markovsky), `mn@ecs.soton.ac.uk` (Mahesan Niranjan)

is less than the perturbation size $\varepsilon$. Therefore, provided that the size $\varepsilon$ of the perturbation $E$ is known, the distance measure $d(X,r)$, for $r = 1,2,\ldots$, can be used to estimate the rank of the unperturbed matrix as follows

$$\widehat{r} = \arg\min\{\, r \mid d(X,r) < \varepsilon \,\}.$$

It is well known that problem (1) has analytic solution in terms of the singular values $\sigma_1,\ldots,\sigma_{\min(m,n)}$ of $X$

$$d(X,r_0) := \sqrt{\sigma_{r_0+1}^2 + \cdots + \sigma_{\min(m,n)}^2},$$

and therefore the rank of $X_0$ can be estimated from the decay of the singular values of $X$ (find the largest singular value that is sufficiently small compared to the perturbation size $\varepsilon$). This is the standard way for rank estimation in numerical linear algebra, where the estimate $\widehat{r}$ is called *numerical rank of $X$*. The question occurs:

> Given a perturbed matrix $X := X_0 + E$, is the numerical rank of $X$ the "best" estimate for the rank of $X_0$, and if so, in what sense?

The answer to the above question depends on the type of the perturbation $E$. If $E$ is a random matrix with zero mean elements that are normally distributed, independent, and with equal variances, then the estimate $\widehat{X}$, defined by (1) is a maximum likelihood estimator of $X_0$, i.e., it is statistically optimal. If, however, one or more of the above assumptions are not satisfied, $\widehat{X}$ is not optimal and can be improved by modifying problem (1). The objective of this paper is to justify this statement in a particular case when there is prior information about the true matrix $X_0$ in the form of structure in a normalized rank-revealing factorization and the elements of the perturbation $E$ are independent but possibly with different variances.

*Prior knowledge in the form of structure*

In applications often there is prior knowledge about the unperturbed matrix $X_0$ (apart from the basic one that $X_0$ is rank deficient). Whenever available, such prior knowledge is beneficial to use in the computation of the distance measure $d(X,r)$. Using the prior knowledge amounts to modification of problem (1). For example, common prior information in image and text classification is nonnegativity of the elements of $X_0$, see [7]. In this case, we require the approximation $\widehat{X}$ to be nonnegative and in order to achieve this, we impose nonnegativity of the estimate $\widehat{X}$ as an extra constraint in (1). Similarly, in signal processing and system theory the matrix $X_0$ is Hankel or Toeplitz structured [11] and the relevant modification of (1) is to constrain $\widehat{X}$ to have the same structure. In chemometrics, the measurement errors $e_{ij}$ may have different variances $\sigma^2 v_{ij}$, which are known (up to a scaling factor) from the measurement setup or from repeated experiments, see [17, 9]. Such prior information amounts to changing the cost function $\|X-\widehat{X}\|_{\mathrm{F}}$ to an element-wise weighted norm of the error matrix $X-\widehat{X}$

$$\|X-\widehat{X}\|_W := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij}(x_{ij}-\widehat{x}_{ij})^2},$$

where the elements $w_{ij}$ of the weight matrix $W \in \mathbb{R}^{m\times n}$ are the inverses of the error variances $v_{ij}$. In general, either the addition of constraints on $\widehat{X}$ or the replacement of the Frobenius norm with a weighted norm, renders the modified distance problem (1) difficult to solve. A globally optimal solution can no longer be given in terms of the singular values of $X$ and the resulting optimization problem is nonconvex.

Most of the approaches to compute low-rank approximation with weighted cost function and constraints are based on local optimization methods, and fall into one of two main classes:

1. methods based on the variable projections [3], and
2. methods based on the alternating projections.

The alternating projections type algorithms are globally convergent with linear local convergence rate [6, 4]. The variable projections type algorithms, when properly implemented, are globally convergent with superlinear local convergence rate. This implies that when the initial approximation is sufficiently close to a local minimum, the variable projections type algorithms are faster than the alternating projections type algorithms. Numerical results [13], however, suggest that in practice, when the initial approximation is "far" from a local minimum, the two approaches are comparable in efficiency.

In this paper, we use the alternating projections approach, because of the easier to modify it for constrained optimization problems. We note that certain constrained problems can be treated also using a modification of the variable projections, see [14, Chapter 8]. Solving constrained low-rank approximation problems via the variable projections approach will be pursued elsewhere.

*Application in bioinformatics*

*Microarray data analysis*

One motivation for our interest in structured factorization comes from the analysis of high throughput gene expression data, measured with microarray technology, where the interest is in inferring the regulatory processes. Expression data correspond to the average concentrations of messenger RNA molecules in a sample of cells. While most work in the use of microarrays deals with static systems, such as profiles of patients with and without a particular disease, there has been growing interest in the modelling of time-course data, either in studying the response of an organism to a particular type of environmental stress, or in steady state dynamical behaviors such as cell-cycle regulation. What motivates us is the last of these, where the expression data is in the form of a matrix $X$, row-wise indexed by the genes in the genome of the organism, and column-wise indexed by time. Typical datasets, where periodic behavior has been the subject of interest, contain two or three periods of the phenomenon, following some experimental method to synchronize the cells in a colony. The classic dataset in this domain is the cell-cycle experiments conducted by Spellman et al. [15] where four different methods were used to synchronize cells, followed by measurements of the expression profiles at a number of equally spaced intervals over two periods. A more recent study [16] focused on the regulation of yeast metabolic cycle, and included three periods of cyclic behavior. Other studies of this nature include the monitoring of Circadean rhythm in plants and cultured.

Such dynamical behavior in which a large number of genes in the organism can be shown to be expressed periodically is regulated by a much smaller set of regulatory proteins, known as transcription factors. In the regulation of yeast cell cycle behavior for example, Fourier transform based estimations detect about 600 genes to be regulated in a cyclic manner. However the number of regulators known to control this behavior is less than 30. An important aspect of biology that justifies the need for model based inference in this topic is the fact that the measured mRNA profiles of the regulators is not an accurate reflection of their regulatory activities. Part of the reason for this is that transcription factors are usually found in low abundances in cells, and hence their measurements are subject to noise. Further, as a result of phenomena known as post-transcriptional and post-translational regulation messenger RNA levels do not correlate well with protein levels. This is particularly true for regulatory proteins, and there is evidence that a significant fraction of cell cycle regulating transcription factors in yeast, for example, are subject to post-transcriptional regulation.

Matrix factorization techniques have been used in the analysis of microarray data in a number of studies [1, 5, 8, 12]. Alter and Golub [1] seek a principal component projection to visualize high dimensional gene expression data and show that some known biological aspects of the data are visible in a two dimensional subspace defined by the first two principal components. The *network component analysis* model uses a factorization of the form $X = CP$, where $C$ the connectivity matrix is rich in structure from prior knowledge of which transcription factors bind to the upstream regions of which genes. Sanguinetti et al. [12] study a variant of this model in a probabilistic state space formulation and estimate parameters using Bayesian methods. Chang et al. [2] has developed a fast computational algorithm to estimate what is called a network component analysis model.

*Formulation as an approximate low-rank factorization with structured factors*

The measurements of a microarray experiment are collected in an $m \times n$ real matrix $X$—rows correspond to genes and columns correspond to time instants. The element $x_{ij}$ is the *expression level* of the $i$th gene at the $j$th moment of time. The rank-$r$ of $X$ is equal to the number of *transcription factors* that regulate the gene expression levels. In a rank revealing factorization $X = CP$, the $j$th column of $P$ is a vector of intensities of the transcription factors at time $j$, and the $i$th row of $C$ is a vector of sensitivities of the $i$th gene to the transcription factors. For example, $c_{ij}$ equal to zero means that the $j$th transcription factor does not regulate the $i$th gene.

An important problem in bioinformatics is to discover what transcription factors regulate a particular gene and what the time evaluation of the transcription factor activities are. This problem amounts to computing an (approximate) factorization $CP$ of the gene expression level versus time matrix $X$. The need of approximation comes from: 1) inability to account for all relevant transcription factors (therefore accounting only for a few dominant ones), and 2) measurement errors occurring in the collection of the data.

Often it is known a priori that certain transcription factors do not regulate certain genes. This implies that certain elements of the sensitivity matrix $C$ are known to be zeros. In addition, the transcription factor activities are modeled to be nonnegative, smooth, and periodic functions of time. Where transcription factors down regulate a gene, the elements of $C$ have to be negative to account for this. The constraints (11–14) in the considered estimation problem (9) (see Section 2) encapsulate this prior knowledge.

A factorization $X = CP$ is nonunique; for any $r \times r$ nonsingular matrix $T$, we obtain a new factorization $X = \widetilde{C}\widetilde{P}$, where $\widetilde{C} := CT^{-1}$ and $\widetilde{P} = TP$. Obviously, this imposes a problem in estimating transcription factor intensities and gene sensitivities from data. In order to resolve the nonuniqueness problem, we assume that $r$ genes are known to be regulated by single transcription factors that are different. Moreover, the sensitivities of these genes to the corresponding transcription factors are normalized to ones. The assumption implies that after reordering of the genes, the sensitivity matrix has the form $C = \left[\begin{smallmatrix} I_r \\ C' \end{smallmatrix}\right]$, where $I_r$ is the $r \times r$ identity matrix. This assumption corresponds to constraint (10) in the estimation problem (9).

In this paper we present an algorithm for approximate low-rank factorization with structured factors and test its performance on synthetic data. A paper on its application to yeast metabolic cycle regulation will be presented elsewhere.

*Notation*

| | |
|---|---|
| $:= \quad (=:)$ | left (right) hand side is defined by the right (left) hand side |
| $A \geq 0$ | matrix with element-wise nonnegative elements, i.e., $a_{ij} \geq 0$ |
| $\lVert A \rVert_{\mathrm{F}} := \sqrt{\sum_{ij} a_{ij}^2}$ | Frobenius norm of $A \in \mathbb{R}^{m \times n}$ |
| $\lVert A \rVert_W := \sqrt{\sum_{ij} w_{ij} a_{ij}^2}$ | element-wise weighted norm with weight $W \in \mathbb{R}^{m \times n}$, $W \geq 0$ |
| $\operatorname{diag}(\cdot)$ | form a diagonal matrix out of a vector $\operatorname{diag}(w) := \begin{bmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{bmatrix}$ |
| $\operatorname{vec} : \mathbb{R}^{m \times n} \mapsto \mathbb{R}^{mn}$ | operator vectorizing a matrix column-wise |
| $\operatorname{vec}^{-1} : \mathbb{R}^{mn} \mapsto \mathbb{R}^{m \times n}$ | operator reconstructing the matrix $A$ back from $\operatorname{vec}(A)$ |
| $\otimes$ | Kronecker product $A \otimes B := [a_{ij}B]$ |
| $\mathbf{1}_n = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ | vector with $n$ elements that are all ones |
| $e \sim \mathrm{N}(m_e, V_e)$ | normally distributed random vector with mean $m_e$ and variance $V_e$ |
| selector matrix | an $m \times n$ matrix $S$ zeros/ones elements, such that $S\mathbf{1}_n = \mathbf{1}_m$ |
| difference matrix | $D := \begin{bmatrix} 1 & & & -1 \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}$ |

## 2. Statistical model and maximum likelihood estimation problem

Consider the errors-in-variables model

$$X = X_0 + E, \quad \text{where} \quad X_0 = C_0 P_0, \quad C_0 \in \mathbb{R}^{m \times r}, \quad P_0 \in \mathbb{R}^{r \times n}, \quad \text{with} \quad r < \min(m,n)$$
$$\text{and} \qquad \operatorname{vec}(E) \sim \mathrm{N}\big(0, \sigma^2 \operatorname{diag}(v)\big). \tag{2}$$

The *true data matrix* $X_0$ has rank $r$ and the measurement errors $e_{ij}$ are zero mean, normal, and uncorrelated, with covariance $\sigma^2 v_{i+m(j-1)}$. The vector $v \in \mathbb{R}^{mn}$ specifies the element-wise variances of the measurement error matrix $E$ up to an unknown factor $\sigma^2$.

In order to make the parameters $C_0$ and $P_0$ unique, we impose the normalization constraint

$$C_0 = \begin{bmatrix} I_r \\ C_0' \end{bmatrix}. \tag{3}$$

In addition, the block $C_0'$ of $C_0$ has elements (specified by a selector matrix $S$) equal to zero

$$S \operatorname{vec}(C_0') = 0. \tag{4}$$

The parameter $P_0$ is periodic with a period $l \in \mathbb{N}$

$$P_0 = \mathbf{1}_l^\top \otimes P_0', \tag{5}$$

nonnegative

$$P_0' \geq 0, \tag{6}$$

and with smooth rows in the sense that

$$\|P_0' D\|_{\mathrm{F}}^2 \leq d, \tag{7}$$

where $d > 0$ is a smoothness parameter.

Define the $m \times n$ matrix

$$W := \operatorname{vec}^{-1}\big(v_1^{-1/2}, \dots, v_{mn}^{-1/2}\big). \tag{8}$$

The maximum likelihood estimator for the parameters $C_0$ and $P_0$ in (2) under assumptions (3–7), with known parameters $r$, $v$, $S$, and $d$, is given by the following optimization problem:

$$\begin{array}{llll}
\text{minimize} \quad \text{over } C', P', \widehat{X} & \|X - \widehat{X}\|_W^2 & \text{(cost function)} & (9) \\
\text{subject to} & \widehat{X} = CP & \text{(rank constraint)} & \\
& C = \begin{bmatrix} I_r \\ C' \end{bmatrix} & \text{(normalization of } C) & (10) \\
& S \operatorname{vec}(C') = 0 & \text{(zero elements of } C') & (11) \\
& P = \mathbf{1}_l^\top \otimes P' & \text{(periodicity of } P) & (12) \\
& P' \geq 0 & \text{(nonnegativity of } P) & (13) \\
& \|P' D\|_{\mathrm{F}}^2 \leq d & \text{(smoothness of } P) & (14)
\end{array}$$

The rank and measurement errors assumptions in the model (2) imply the weighted low-rank approximation nature of the estimation problem (9–14) with weight matrix given by (8). Furthermore, the assumptions (3–7) about the true data matrix $X_0$ correspond to the constraints (10–14) in the estimation problem.

## 3. Computational algorithm

### 3.1. Algorithm

The alternating projections algorithm, see Algorithm 1, is based on the observation that the cost function (9) is quadratic and the constraints (10–14) are linear in either

**Algorithm 1** Alternating projections algorithm for solving problem (9–14).

```
• Find an initial approximation $(C'^{(0)}, P'^{(0)})$.
• For $k = 0, 1, \ldots$ till convergence do
    1. $C'^{(k+1)} := \arg\min_{C'} \|X - CP\|_W^2$   subject to   (10–11) with $P' = P'^{(k)}$
    2. $P'^{(k+1)} := \arg\min_{P'} \|X - CP\|_W^2$   subject to   (12–14) with $C' = C'^{(k+1)}$
```

$C$ or $P$. Therefore, for a fixed value of $C$, (9–14) is a nonnegativity constrained least squares problem in $P$ and vice versa, for a fixed value of $P$, (9–14) is a constrained least squares problem in $C$. These problems correspond to, respectively, steps 1 and 2 of the algorithm. Geometrically they are projections. In the unweighted (i.e., $W = \mathbf{1}_m \mathbf{1}_n^\top$) and unconstrained case, the problem on step 1 is the orthogonal projection $X(PP^\top)^{-1}P^\top$ of $X$ on the span of the rows of $P$, and problem on step 2 is the orthogonal projection $(C^\top C)^{-1}C^\top X$ of $X$ on the span of the column of $C$. The algorithm iterates the two projections, thus its name—alternating projections.

*Note* 1 (Rank deficient factors $C$ and $P$). If the factor $P$ is rank deficient, the indicated inverse in the computation of the projected matrix $C^*$ does not exist. (This may happen when the rank of the approximation $\widehat{X}$ if less than $r$.) The projection $C^*$, however, is still well defined by the optimization problem on step 1 of the algorithm and can be computed in closed form by replacing the inverse with the pseudo inverses. The same is true when the factor $P$ is rank deficient.

Two special cases of the estimation problem of Section 2 are studied and alternating projections type algorithms are proposed.

- In the weighted and unconstrained case, Algorithm 1 is equivalent to the MLPCA algorithm of [17].
- In the weighted case with nonnegativity constraint (13), Algorithm 1 is equivalent to the modified MLPCA algorithm of [18].

In Appendix A we describe the implementation of Algorithm 1 for the general case of inhomogeneous weights and constraints (10–14). Next, we state convergence properties of Algorithm 1.

### *3.2. Convergence properties*

**Theorem 2.** *Algorithm 1 is globally and monotonically convergent in the $\|\cdot\|_W$ norm, i.e., if $\widehat{X}^{(k)} := C^{(k)}P^{(k)}$ is the approximation on the kth step of the algorithm, then*

$$f(k) := \|X - \widehat{X}^{(k)}\|_W^2 \to f^*, \qquad \text{as } k \to \infty. \tag{15}$$

*Assuming that there exists a solution to the problem (9–14) and any (locally optimal) solution is unique (i.e., it is a strict minimum), the sequences $\widehat{X}^{(k)}$, $C^{(k)}$, and $P^{(k)}$ converge element-wise, i.e.,*

$$\widehat{X}^{(k)} \to X^*, \quad C^{(k)} \to C^*, \quad \text{and} \quad P^{(k)} \to P^*, \qquad \text{as } k \to \infty, \tag{16}$$

*where $X^* := C^* P^*$ is a (locally optimal) solution of (9–14).*

*Proof.* First, we show that the sequence $\widehat{X}^{(k)}$, for $k = 1, 2, \ldots$, converges monotonically in the $\|\cdot\|_W$ norm. On each iteration, Algorithm 1 solves two optimization problems (steps 1 and 2), which cost function and constraints coincide with the ones of problem (9–14). Therefore, the cost function $\|X - \widehat{X}^{(k)}\|_W^2$ is monotonically nonincreasing. The cost function is bounded from below, so that the sequence $\|X - \widehat{X}^{(k)}\|_W^2$, for $k = 1, 2, \ldots$, is convergent. This proves (15).

Although, $\widehat{X}^{(k)}$ converges in norm, it may not converge element-wise. A sufficient condition for element-wise convergence is that the underlying optimization problem has a solution and it is unique [4, Theorem 5]. The element-wise convergence of $\widehat{X}^{(k)}$ and the uniqueness (due to the normalization condition (3)) of the factors $C^{(k)}$ and $P^{(k)}$, given $\widehat{X}^{(k)}$, implies element-wise convergence of the factor sequences $C^{(k)}$ and $P^{(k)}$ as well. This proves (16).

In order to show that the algorithm convergence to a minimum point of (9–14), we need to verify that the first order optimality conditions for (9–14) are satisfied at a cluster point of the algorithm. The algorithm converges to a cluster point if and only if the union of the first order optimality conditions for the problems on steps 1 and 2 are satisfied. Then

$$P'^{(k-1)} = P'^{(k)} =: P'^* \qquad \text{and} \qquad C'^{(k-1)} = C'^{(k)} =: C'^*.$$

From the above conditions for a stationary point and the Lagrangians of the problems of steps 1 and 2 and (9–14), it is easy to see that the union of the first order optimality conditions for the problems on steps 1 and 2 coincides with the first order optimality conditions of (9–14). □

## 4. Simulation results

In this section, we show empirically that exploiting prior knowledge ((8) and assumptions (3–7)) improves the performance of the estimator. The data matrix $X$ is generated according to the errors-in-variables model (2) with parameters $m = 100$, $n = 6$, and $r = 2$. The true low-rank matrix $X_0 = C_0 P_0$ is random and the parameters $C_0$ and $P_0$ are normalized according to assumption (3) (so that they are unique). For the purpose of validating the algorithm, the element $c_{0,mn}$ is set to zero but this prior knowledge is not used in the parameter estimation.

The estimation algorithm is applied on $N = 100$ independent noise realizations of the data $X$. The estimated parameters on the $i$th repetition are denoted by $\widehat{C}^i$, $\widehat{P}^i$ and $\widehat{X}^i := \widehat{C}^i \widehat{P}^i$. The performance of the estimator is measured by the following average relative estimation errors:

$$e_X = \frac{1}{N} \sum_{i=1}^{N} \frac{\|X_0 - \widehat{X}^i\|_F^2}{\|X_0\|_F^2}, \qquad e_C = \frac{1}{N} \sum_{i=1}^{N} \frac{\|C_0 - \widehat{C}^i\|_F^2}{\|C_0\|_F^2}, \qquad e_P = \frac{1}{N} \sum_{i=1}^{N} \frac{\|P_0 - \widehat{P}^i\|_F^2}{\|P_0\|_F^2},$$

$$\text{and} \quad e_z = \frac{1}{N} \sum_{i=1}^{N} |\widehat{c}_{mn}^i|.$$

For comparison the estimation errors are reported for the low-rank approximation algorithm, using only the normalization constraint (3), as well as for the proposed algorithm, exploiting the available prior knowledge. The difference between the two estimation errors is an indication of how important is the prior knowledge in the estimation.

Lack of prior knowledge is reflected by specific choice of the simulation parameters as follows:

| | | |
|---|---|---|
| homogeneous errors | ↔ | $W = \texttt{ones}(\texttt{m},\texttt{n})$ |
| no periodicity | ↔ | $l = 1$ |
| no zeros in $C'$ | ↔ | $S = []$ |
| no sign constraint on $P'$ | ↔ | $\texttt{nonneg} = 0$ |

We perform the following experiments:

1. $W = \texttt{rand}(\texttt{m},\texttt{n})$, $l = 1$, $S = []$, $\texttt{nonneg} = 0$
2. $W = \texttt{ones}(\texttt{m},\texttt{n})$, $l = 3$, $S = []$, $\texttt{nonneg} = 0$
3. $W = \texttt{ones}(\texttt{m},\texttt{n})$, $l = 1$, $S \neq []$, $\texttt{nonneg} = 0$
4. $W = \texttt{ones}(\texttt{m},\texttt{n})$, $l = 1$, $S = []$, $\texttt{nonneg} = 1$
5. $W = \texttt{rand}(\texttt{m},\texttt{n})$, $l = 3$, $S \neq []$, $\texttt{nonneg} = 1$

which test individually the effect of (8), assumptions (4), (5), (6), and their combined effect on the estimation error. Figures 1–5 show the average relative estimation errors (solid blue line is the estimator that exploits prior knowledge and dashed red line is the estimator that does not exploit prior knowledge) versus the measurement noise standard deviation $\sigma$, for the five experiments. The vertical bars on the plots visualize the standard deviation of the estimates. The results indicate that main factors for the improved performance of the estimator are:

1. assumption (5) — known zeros in the $C_0'$ and
2. (8) — known covariance structure of the measurement noise.

MATLAB files reproducing the numerical results and figures presented in the paper are available from: `http://users.ecs.soton.ac.uk/im/factorize.tar`

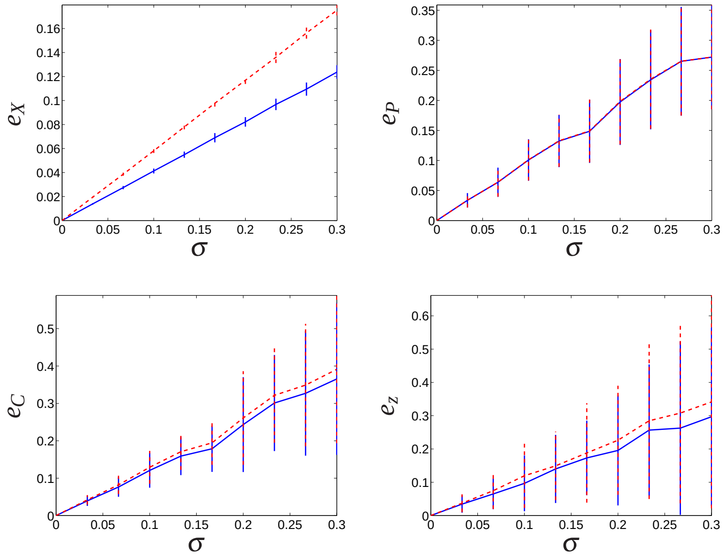

Figure 1: Effect of weighting (solid blue line — exploiting prior knowledge, dashed red line — without exploiting prior knowledge, vertical bars — standard deviations).

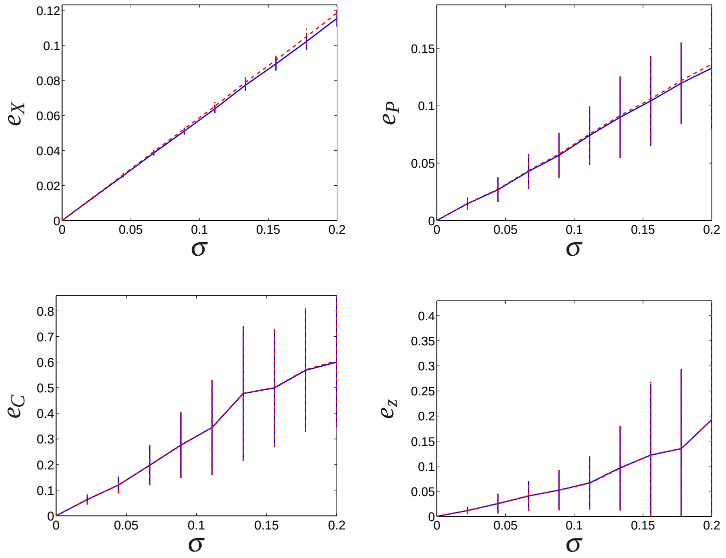

Figure 2: Effect of periodicity of $P$ (solid blue line — exploiting prior knowledge, dashed red line — without exploiting prior knowledge, vertical bars — standard deviations).

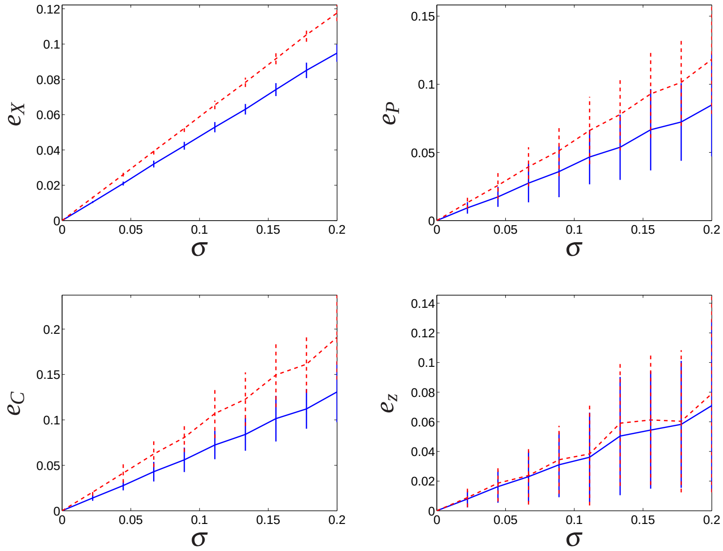

Figure 3: Effect of zero elements in $C$ (solid blue line — exploiting prior knowledge, dashed red line — without exploiting prior knowledge, vertical bars — standard deviations).

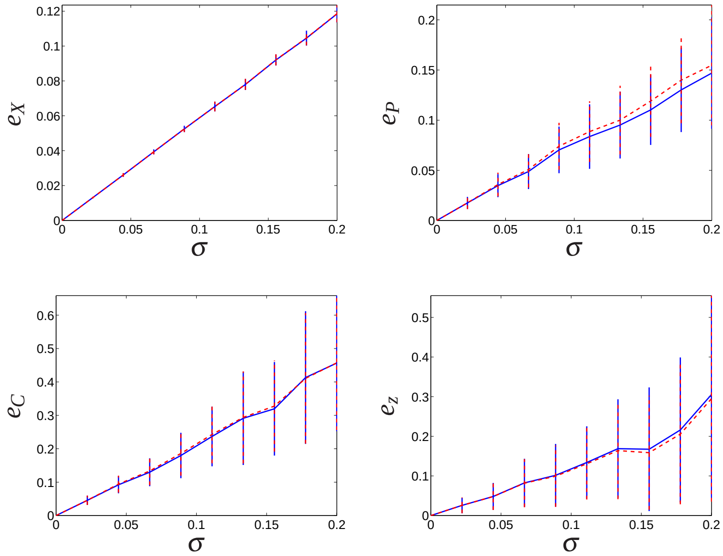

Figure 4: Effect of nonnegativity of $P$ (solid blue line — exploiting prior knowledge, dashed red line — without exploiting prior knowledge, vertical bars — standard deviations).

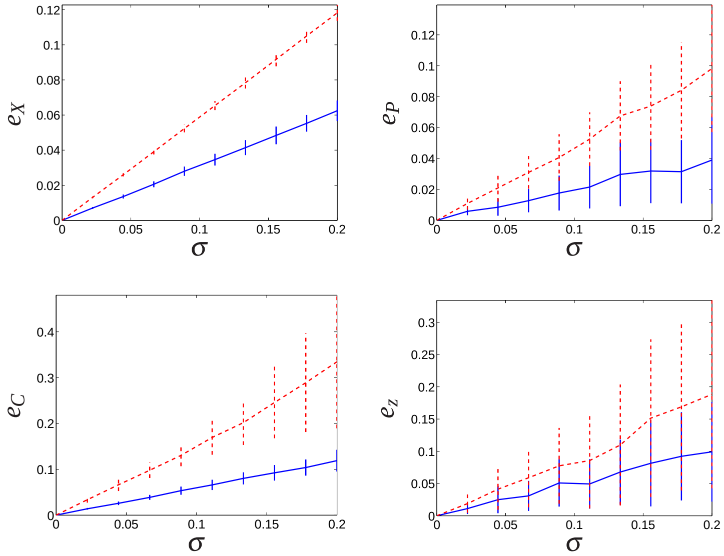

Figure 5: Effect of weighting, periodicity, and nonnegativity of $P$, and zero elements in $C$ (solid blue line — exploiting prior knowledge, dashed red line — without exploiting prior knowledge, vertical bars — standard deviations).

## 5. Conclusions

The low-rank approximation accuracy improves when there is prior knowledge about the to-be-estimation data matrix or the perturbations and this prior knowledge is used in the approximation problem. Using the prior knowledge changes the basic low-rank approximation problem in the Frobenius norm to a weighted constrained low-rank approximation problem. Unfortunately, the latter problem is, in general, difficult nonconvex optimization problem, while the former is solvable in terms of the singular value decomposition. We adopted a solution approach for the weighted constrained low-rank approximation problem that is based on an alternating projection algorithm. The alternating projection algorithm is globally convergent to a local solution, the convergence is monotonic, and has linear local rate. An interesting research question for future research is to use first and second derivative information in order to speed up the convergence (e.g., achieve superlinear convergence rate). In the specific estimation problem considered in the paper, the simulation results suggest that the improvement in the estimation accuracy is mainly due to known zeros in a factor of the normalized rank revealing factorization and the known covariance structure of the measurement noise.

## References

[1] Alter, O., Golub, G. H., 2006. Singular value decomposition of genome-scale mRNA lengths distribution reveals asymmetry in RNA gel electrophoresis band

broadening. Proceedings of the National Academy of Sciences 103, 11828–11833.

[2] Chang, C., Ding, Z., Hung, Y., Fung, P., 2008. Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data. Bioinformatics 24 (11), 1349–1358.

[3] Golub, G., Pereyra, V., 2003. Separable nonlinear least squares: the variable projection method and its applications. Institute of Physics, Inverse Problems 19, 1–26.

[4] Kiers, H. A., 2002. Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems. Comput. Statist. Data Anal. 41, 157–170.

[5] Kim, H., Park, H., 2007. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics 23, 1495–1502.

[6] Krijnen, W. P., 2006. Convergence of the sequence of parameters generated by alternating least squares algorithms. Comput. Statist. Data Anal. 51, 481–489.

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

[8] Liao, J., Boscolo, R., Yang, Y., Tran, L., Sabatti, C., Roychowdhury, V., 2003. Network component analysis: Reconstruction of regulatory signals in biological systems. Proceedings of the National Academy of Sciences of the United States of America 100 (26), 15522–15527.

[9] Markovsky, I., Rastello, M.-L., Premoli, A., Kukush, A., Van Huffel, S., 2005. The element-wise weighted total least squares problem. Comput. Statist. Data Anal. 50 (1), 181–209.

[10] Markovsky, I., Van Huffel, S., 2007. Left vs right representations for solving weighted low rank approximation problems. Linear Algebra Appl. 422, 540–552.

[11] Markovsky, I., Van Huffel, S., Pintelon, R., 2005. Block-Toeplitz/Hankel structured total least squares. SIAM J. Matrix Anal. Appl. 26 (4), 1083–1099.

[12] Sanguinetti, G., Rattray, M., Lawrence, N., 2006. A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription. Bioinformatics 22 (14), 1753–1759.

[13] Schuermans, M., Markovsky, I., Wentzell, P., Van Huffel, S., 2005. On the equivalence between total least squares and maximum likelihood PCA. Analytica Chimica Acta 544, 254–267.

[14] Sima, D., 2006. Regularization techniques in model fitting and parameter estimation. Ph.D. thesis, Faculty of Engineering, K.U.Leuven.

[15] Spellman, P., et al., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell 9, 3273–3297.

[16] Tu, B., Kudlicki, A., Rowicka, M., McKnight, S., 2005. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. Science 310 (5751), 1152–1158.

[17] Wentzell, P., Andrews, D., Hamilton, D., Faber, K., Kowalski, B., 1997. Maximum likelihood principle component analysis. J. Chemometrics 11, 339–366.

[18] Wentzell, P., Karakach, T., Roy, S., Martinez, M., Allen, C., Werner-Washburne, M., 2006. Multivariate curve resolution of time course microarray data. BMC Bioinformatics 7, 343.

## A. Implementation of Algorithm 1

*Initial approximation.* For initial approximation $(C'^{(0)}, P'^{(0)})$ we choose the normalized factors of a rank revealing factorization of the solution $\widehat{X}$ of (1). Let $X = U\Sigma V^\top$ be the singular value decomposition of $X$ and define the partitioning

$$U =: \begin{array}{c} \begin{array}{cc} r & m-r \end{array} \\ \begin{bmatrix} U_1 & U_2 \end{bmatrix} \end{array}, \quad \Sigma =: \begin{array}{cc} \begin{array}{cc} r & n-r \end{array} & \\ \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} & \begin{array}{c} r \\ m-r \end{array} \end{array}, \quad V =: \begin{array}{c} \begin{array}{cc} r & n-r \end{array} \\ \begin{bmatrix} V_1 & V_2 \end{bmatrix} \end{array}.$$

Furthermore, let $\begin{bmatrix} U_{11} \\ U_{21} \end{bmatrix} := U$, with $U_{11} \in \mathbb{R}^{r \times r}$. Then

$$C'^{(0)} := U_{21} U_{11}^{-1} \qquad \text{and} \qquad P^{(0)} := U_{11} \Sigma V^\top$$

define the Frobenius-norm optimal unweighted and unconstrained low-rank approximation

$$\widehat{X}^{(0)} := \begin{bmatrix} I \\ C'^{(0)} \end{bmatrix} P^{(0)}.$$

More sophisticated choices for the initial approximation that take into account the weight matrix $W$ are described in [10].

*Separable least squares problem for C.* In the weighted case, the projection on step 1 of the algorithm is computed separately for each row $c_i$ of $C$. Let $x_i$ be the $i$th row of $X$ and $w_i$ be the $i$th row of $W$. The problem

$$\text{minimize}_C \quad \|X - CP\|_W^2 \quad \text{subject to} \quad (10\text{–}11)$$

is equivalent to the problem

$$\text{minimize}_{c_i} \quad \|(x_i - c_i P)\,\text{diag}(w_i)\|_2^2 \quad \text{subject to} \quad (10\text{–}11), \qquad \text{for } i = 1, \ldots, m. \quad (17)$$

The projection on step 2 of the algorithm is not separable due to constraint (14) (see below).

*Taking into account constraint (10).* Since the first $r$ rows of $C$ are fixed, we do not solve (17) for $i = 1, \dots, r$, but define

$$c_i := e_i^\top, \qquad \text{for} \quad i = 1, \dots, r,$$

where $e_i$ is the $i$th unit vector (the $i$th column of the identity matrix $I_r$).

*Taking into account constraint (11).* Let $S_i$ be a selector matrix for the zeros in the $i$th row of $C$

$$S \operatorname{vec}(C') = 0 \quad \iff \quad c_i S_i = 0, \text{ for } i = r+1, \dots, m.$$

(If there are no zeros in the $i$th row, then $S_i$ is skipped.) The $i$th problem in (17) becomes

$$\text{minimize}_{c_i} \quad \|(x_i - c_i P) \operatorname{diag}(w_i)\|_2^2 \quad \text{subject to} \quad c_i S_i = 0. \tag{18}$$

Let the rows of the matrix $N_i$ form a basis for the left null space of $S_i$. Then $c_i S_i = 0$ if and only if $c_i = z_i N_i$, for certain $z_i$, and problem (18) becomes

$$\text{minimize}_{z_i} \quad \|(x_i - z_i N_i P) \operatorname{diag}(w_i)\|_2^2.$$

Therefore, the solution of (17) is

$$c_i^* = x_i P^\top N_i^\top (N_i P P^\top N_i^\top)^{-1} N_i.$$

*Note* 3. It is not necessary to explicitly construct the matrices $S_i$ and compute basis $N_i$ for their left null spaces. Since $S_i$ is a selector matrix, it is a submatrix of the identity matrix $I_r$. The rows of the complementary submatrix of $I_r$ form a basis for the left null space of $S_i$. This particular matrix $N_i$ is also a selector matrix, so that the product $N_i P$ need not be computed explicitly.

*Taking into account constraint (12).* We have,

$$\begin{aligned} X - CP = X - C(\mathbf{1}_l^\top \otimes P') &= \begin{bmatrix} X_1 & \cdots & X_l \end{bmatrix} - C \begin{bmatrix} P' & \cdots & P' \end{bmatrix} \\ &= \begin{bmatrix} X_1 \\ \vdots \\ X_l \end{bmatrix} - \begin{bmatrix} C \\ \vdots \\ C \end{bmatrix} P' =: X' - \underbrace{(\mathbf{1}_l \otimes C)}_{C'} P' = X' - C'P'. \end{aligned}$$

Let $W' := \begin{bmatrix} W_1 \\ \vdots \\ W_l \end{bmatrix}$, where $W =: \begin{bmatrix} W_1 & \cdots & W_2 \end{bmatrix}$. Then the problem

$$\text{minimize}_P \, \|X - CP\|_W^2 \quad \text{subject to} \quad (12\text{–}14)$$

is equivalent to the problem

$$\text{minimize}_{P'} \, \|X' - C'P'\|_{W'}^2 \quad \text{subject to} \quad (13\text{–}14).$$

*Taking into account constraint (13).* Adding the nonnegativity constraint changes the least squares problem to a nonnegative least squares problem, which is a standard convex optimization problem for which robust and efficient methods and software exist.

*Taking into account constraint (14).* The problem

$$\text{minimize}_P \quad \|X - CP\|_W^2 \quad \text{subject to} \quad \|PD\|_F^2 \leq \delta$$

is equivalent to a Tychonov regularized least squares problem

$$\text{minimize}_P \quad \|X - CP\|_W^2 + \gamma \|PD\|_F^2$$

for certain regularization parameter $\gamma$. The latter problem is equivalent to the standard least squares problem

$$\text{minimize}_p \quad \left\| \begin{bmatrix} \operatorname{diag}\big(\operatorname{vec}(W)\big) & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x - \mathscr{C} p \\ \sqrt{\gamma} \mathscr{D} p \end{bmatrix} \right\|_2^2$$

where $p = \operatorname{vec} P$, $\mathscr{C} := I \otimes C$, and $\mathscr{D} := D^\top \otimes I$.

*Stopping criteria.* The iteration is terminated when the following stopping criteria are satisfied

- $\|C^{(k+1)}P^{(k+1)} - C^{(k)}P^{(k)}\|_W / \|C^{(k+1)}P^{(k+1)}\|_W < \varepsilon_X$,
- $\|(C^{(k+1)} - C^{(k)})P^{(k+1)}\|_W / \|C^{(k+1)}P^{(k+1)}\|_W < \varepsilon_C$, and
- $\|C^{(k+1)}(P^{(k+1)} - P^{(k)})\|_W / \|C^{(k+1)}P^{(k+1)}\|_W < \varepsilon_P$.

Here $\varepsilon_X$, $\varepsilon_P$, and $\varepsilon_C$ are user defined relative convergence tolerances for $X$, $P$, and $C$, respectively.