Approximation of non–negative integer–valued matrices with application to Hungarian mortality data

Márton Ispány, György Michaletzky, Jenő Reiczigel, Gábor Tusnády, Paula Tusnády, and Katalin Varga

***Abstract*— Singular valued decomposition (SVD) is a commonly applied technique for dimensionality reduction. SVD implicitely minimizes an unweighted sum of squares which may be inappropriate in several practical applications. This paper gives generalizations of SVD to other loss functions, e.g., weighted Frobenius distance and logistic loss, that are better suited to the data. We describe algorithms for minimizing these loss functions, and give an application to Hungarian mortality data.**

***Index Terms*— Singular valued decomposition; Weighted low–rank approximation; Fisher scoring; Count data; Non–negative matrix approximation; Maximum likelihood; Logistic loss;**

I. INTRODUCTION

Low dimensional data representations are primordial to numerous applications in statistics, machine learning, signal processing, and bioinformatics. Singular valued decomposition (SVD) suggested by Eckart and Young [8] is one of the most widely used methods for dimensionality reduction. SVD expresses a rectangular matrix via an additive combination of the dyadic (outer) products of dual right and left eigenvectors, see Chapter 4 in [21]. Partial sums of these dyadic products provide optimal solution of the least squares problem of matrices in various norms under low–rank condition, and the error of the approximation can be expressed by the appropriate singular values of the data matrix. Moreover, SVD plays a crucial role in several problems of linear algebra and numerical optimization, see [4] and [21]. Examples of applications can be found in pattern recognition ([1], [20]), cluster analysis ([7]), multidimensional scaling ([17]), correspondence analysis ([3]), data visualization ([6]), time series analysis ([12]), and collaborative filtering ([19]).

Recently, there has been considerable interest in factorization or approximation of matrices which satisfy some additional conditions or constraints, e.g., weighting, nonnegativity, or discreteness. On the one hand, often we obtain heteroscedastic low–rank approximation problem since we have some additional knowledge on the precision of the entries of the data matrix. On the other hand, often the data to be analyzed is nonnnegative which follows from physical realities. The discrete nature of certain data sets may come from the need to count events, objects or individuals. Classical tools like principal component analysis (PCA) and singular valued decomposition (SVD) cannot garantee to maintain these requirements and properties. In this paper we propose a few generalizations of the low–rank matrix approximation and SVD that are better suited in the above mentioned situations, and describe numerical algorithms for computing the approximate matrix efficiently.

In the first part of the paper we consider the low–rank matrix approximation by weighted least squares. This problem has payed relatively little attention up to now. In statistics Wold and Lyttkens [23], later Gabriel and Zamir [10] investigated the weighted case. The former suggested the nonlinear iterative partial least squares (NIPALS) algorithm, while the latter applied the criss–cross multiple regression. In machine learning Srebro and Jaakkola [18] provided a simple EM algorithm for minimizing the weighted sum of squares. Lu et al. [13] suggested alternating optimization method based on combined rank–one approximations. In this paper we propose a new algorithm which is based on the second order Taylor expansion of the objective function. Moreover, our algorithm is a kind of Fisher scoring because the Hessian in the second order term of the expansion is approximated by the Fisher information. Since the local convergence of the algorithm is quadratic, hence it is faster than the above mentioned ones, and thus it is an attractive alternative in numerous practical applications.

In the second part of the paper the low–rank approximation of non–negative integer–valued matrices is considered. As an application we are interested in the approximation of three dimensional contingency tables of size $2 \times I \times J$. These tables arises naturally in the analysis of mortality tables, where the two $I \times J$ dimensional subtables denote the number of individuals who died and survived, respectively, at age i in year j. Several non–negative matrices can be derived from these subtables, e.g., the odds ratio of the dead and the living or the relative frequency of the dead. Then standard non–negative matrix factorization methods, e.g. the Lee–Seung algorithm [14], can be applied, see the review paper of Berry et al. [5]. Instead of these descriptive statistical techniques we propose stochastic generative models for describing contingency tables by low–rank matrices. There are two ways for dimensionality reduction: the inner (indirect) and the outer (direct) factorization. The former one is given by modelling

M. Ispány is with the Department of Information Technology, Faculty of Informatics, University of Debrecen, 4032 Debrecen, Hungary ispany.marton@inf.unideb.hu

Gy. Michaletzky is with Department of Probability Theory and Statistics, Faculty of Science, Eötvös Loránd University, Budapest, Hungary michgy@ludens.elte.hu

J. Reiczigel is with the Department of Biomathematics and Informatics, Faculty of Veterinary Science, Szent István University, Budapest, Hungary Reiczigel.Jeno@aotk.szie.hu

G. Tusnády is with the Alfréd Rényi Mathematical Institute of Hungarian Academy of Sciences, Budapest, Hungary tusnady@renyi.hu

P. Tusnády is with the Hungarian Financial Supervisory Authority, Budapest, Hungary tusnady.paula@pszaf.hu

K. Varga is with the OTP Bank, Budapest, Hungary varga@szit.bme.hu

ISBN 978-963-311-370-7

the $I \times J$ matrix of probabilities of death by Bernoulli distribution where the parameter matrix satisfies a low–rank constraint, see Section III. The latter one is given by approximating the table by a mixture of independent tables. In this case an EM algorithm can be applied similarly to the mixture decomposition of latent class models, see [9]. In the paper we propose a new algorithm, the iteratively reweighted singular valued decomposition (IRSVD), for deriving the inner factorization. This algorithm seems more stable and faster than the EM–type ones in the outer factorization. Moreover, this new factorization technique provides better fit to the data than the standard Lee–Carter model, see Lee and Carter [12] and Baran et al. [2].

The rest of the paper is organized as follows. Section II describes the weighted SVD as a low–rank approximation problem in weighted Frobenius distance. Section III contains stochastic models whose likelihoods lead to the low–rank approximation problem discussed in the paper. Section IV provides the asymptotic theory of these stochastic models. In Section V we propose an efficient algorithm for obtaining weighted dyadic approximation. In Section VI further numerical algorithms are presented in more general cases, particularly, in the logistic low–rank approximation model. Application of these algorithms in the field of mortality study for the Hungarian data is highlighted in Section VII. The proofs are left for the Appendix.

Notation

Let $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{R}_+$ denote the set of positive integers, real numbers and non–negative real numbers, respectively. Matrices are denoted by capital letters ($A, B, \Sigma, \ldots$), vectors by roman lowercase ($u, v, w, \ldots$), and scalars by greek lowercase ($\alpha, \beta, \ldots$). An entry of a matrix A is referred as a_{ij} or $A(i,j)$. The column and row vectors of a matrix $A \in \mathbb{R}^{I\times J}$ are denoted by $a_1, \ldots, a_J$ and $a^1, \ldots, a^I$, respectively. I_d stands for the $d \times d$ identity matrix and δ_{ij} for the Kronecker delta. For any pair of vectors $u, v \in \mathbb{R}^d$ and weight vector $w \in \mathbb{R}^d_+$ the weighted scalar product of u and v is defined by $\langle u, v\rangle_w := \sum_{i=1}^d w_i u_i v_i$. The weighted norm is denoted by $\|u\|_w := \langle u, u\rangle_w^{1/2}$. For any pair of matrices $U, V \in \mathbb{R}^{I\times J}$ and weight matrix $W \in \mathbb{R}_+^{I\times J}$ the weighted Frobenius scalar product of U and V is defined by $\langle U, V\rangle_W := \sum_{i=1}^I \sum_{j=1}^J w_{ij} u_{ij} v_{ij}$. The weighted Frobenius norm $\|A\|_W$ of $A \in \mathbb{R}^{I\times J}$ is defined as the square root of $\langle A, A\rangle_W$. In case of $w_{ij} = 1$ for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$ the Frobenius scalar product and norm are denoted by $\langle\cdot,\cdot\rangle_F$ and $\|\cdot\|_F$, respectively. We denote by $A \circ B$ the entry–wise multiplication (Hadamard product) of matrices A, B, i.e., $(A \circ B)_{ij} := a_{ij} b_{ij}$ for all i and j. For a matrix A its rank is denoted by $\mathsf{rank}(A)$. For matrices $A_{kl} \in \mathbb{R}^{I\times J}$, $k, l = 1, \ldots, K$, denote by $\mathsf{mat}\{A_{kl}, k, l = 1, \ldots, K\}$ the block–matrix of order $IK \times JK$ with entries A_{kl}. The block diagonal matrix of order $IK \times JK$ is denoted by $\mathsf{diag}\{A_k, k = 1, \ldots, K\}$, where $A_k \in \mathbb{R}^{I\times J}$, $k = 1, \ldots, K$. Linear mappings between vector spaces of matrices are denoted by calligraphic capitals ($\mathcal{F}, \mathcal{G}, \ldots$). If $\mathcal{G} : \mathbb{R}^{I\times J} \to \mathbb{R}^{I\times J}$ is a linear mapping then we refer to the entries of its matrix as $\mathcal{G}(k, l; i, j)$. Thus, if $B = \mathcal{G}A$ then $b_{kl} = \sum_{i=1}^I \sum_{j=1}^J \mathcal{G}(k, l; i, j) a_{ij}$ for all k and l.

II. WEIGHTED LOW–RANK APPROXIMATION

Let us briefly describe the weighted SVD, or weighted matrix approximation by cumulative sum of the dyadic (i.e. rank one) matrices. Let D denote a data (or target) matrix of $I \times J$ order, where $I, J \in \mathbb{N}$, with elements d_{ij} of its ith row and jth column. Let us given a corresponding non–negative weight matrix $W \in \mathbb{R}_+^{I\times J}$, and a positive integer K such that $K \le \min\{I, J\}$ for the rank of the approximate matrix. If $w_{ij} = 1$ for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$ then we have the so–called unweighted case. We would like to find matrices $U \in \mathbb{R}^{I\times K}$ and $V \in \mathbb{R}^{J\times K}$ that minimizes the weighted Frobenius distance

$$Q(U, V) := \frac{1}{2}\|D - UV^\top\|_W^2. \tag{1}$$

Denote the column vectors of U and V by $u_1, \ldots, u_K$ and $v_1, \ldots, v_K$, respectively. Thus, in (1), D is approximated by a rank–K matrix $M := UV^\top = \sum_{k=1}^K u_k v_k^\top$. Since any matrix of rank K can be decomposed in such a way, the weighted rank–K approximation is an unconstrained problem over pairs of matrices (U, V) with objective function (1). The pairs (U, V) form an $(I + J)K$–dimensional real vector space, i.e., the dimension of the parameter space for optimization problem (1) is $(I + J)K$. However, this decomposition is not unique because for any invertable $O \in \mathbb{R}^{K\times K}$ the pair (UO, VO^{-1}) yields the same approximate matrix like (U, V), thus $Q(U, V) = Q(UO, VO^{-1})$. These equivalent solutions form a K^2–dimensional manifold suggesting that the proper parameter space has dimension $(I + J - K)K$.

Indeed, taking the SVD of the approximate matrix M we have the following block–matrix representation

$$M = \begin{bmatrix} U_1\Sigma V_1^\top & U_1\Sigma V_2^\top \\ U_2\Sigma V_1^\top & U_2\Sigma V_2^\top \end{bmatrix}, \tag{2}$$

where $\Sigma := \mathsf{diag}\{\sigma_1, \ldots, \sigma_K\}$ with $\sigma_i > 0$, $i = 1, \ldots, K$, singular values of M, and $U_1, V_1 \in \mathbb{R}^{K\times K}$, $U_2 \in \mathbb{R}^{(I-K)\times K}$, $V_2 \in \mathbb{R}^{(J-K)\times K}$ with $U_1^\top U_1 + U_2^\top U_2 = I_K$ and $V_1^\top V_1 + V_2^\top V_2 = I_K$ (see Appendix). Without loss of generality, taking appropriate permutations of rows and columns of the data matrix and the approximate matrix, we may suppose that the matrix $M_1 := U_1\Sigma V_1^\top$ is of rank K. This implies that U_1 and V_1 are invertable matrices. Our parametrization is based on congruence classes of matrices. Recall that two matrices $A, B \in \mathbb{R}^{K\times K}$ are called congruent provided there exists an invertable matrix P such that $B = PAP^\top$. Clearly this is an equivalence relation. The complete description of the congruence classes can be found in [15] by enumerating their representatives. Assume that M_1 is congruent to a representative O with $\|O\|_F = 1$ by $M_1 = POP^\top$. Introduce the matrices $Q = U_2U_1^{-1}P$ and $R = V_2V_1^{-1}P$. Then the matrix M can be expressed as

$$M = \begin{bmatrix} POP^\top & POR^\top \\ QOP^\top & QOR^\top \end{bmatrix} = \begin{bmatrix} P \\ Q \end{bmatrix} O[P^\top R^\top].$$

Thus, if we fix O, we might expect that it is possible to parametrize the neighborhood of M in the space of rank–K matrices by the triplet (P, Q, R), where $P \in \mathbb{R}^{K\times K}$, $Q \in \mathbb{R}^{(I-K)\times K}$ and $R \in \mathbb{R}^{(J-K)\times K}$ confirming that the proper parametrization is $(I + J - K)K$ dimensional. Unfortunately, this is not true in general. If a congruence class is isolated, i.e., there exists a neighborhood of its representative which does not contain any representative of an other congruence class, then this parametrization is proper. For example, we will see later that if $K = 1$ then there are two isolated congruence classes with representatives $O = +1$, and $O = -1$. However, if $K > 1$ then there exist connected congruence classes. In two dimensions a simple example is congruence classes belonging to $O_\alpha := (2 + 2\alpha^2)^{-1}\begin{bmatrix} 0 & 1+\alpha \\ 1-\alpha & 0 \end{bmatrix}$, $\alpha \in \mathbb{R}$. Particularly, the congruence class with representative O_0 is one of Sylvester's classes of symmetric matrices with eigenvalues $+1$ and -1, see Theorem 4.5.8 in [11]. The cases $\alpha \neq 0$ cover more general, skew–symmetric congruent classes. Thus, in these cases we need extra parameters for representatives resulting overparametrization on the surface. However, in connected congruence classes the degrees of freedom in parameter P decrease. For example, if $P_\gamma := \begin{bmatrix} \gamma p_1 & \gamma^{-1} p_2 \\ \gamma p_3 & \gamma^{-1} p_4 \end{bmatrix}$, $\gamma \neq 0$, then $P_\gamma O_\alpha P_\gamma = P_1 O_\alpha P_1$ for all $\gamma \neq 0$ and $\alpha \in \mathbb{R}$, hence the proper dimension of parameter P is 3, i.e., we have again a 4–dimensional parametrization for M_1. A general approach to parametrize the set of matrices of rank K is to partition it into isolated and connected congruence classes and then to parametrize these open sets by the same way. The result of this procedure will be a $(I + J - K)K$ dimensional global parametrization of the set of matrices of rank K. We note that an other, local parametrization of the set of matrices of finite rank can be derived by the implicit function theorem.

Equipped with the above parametrization of the matrices of rank K we may investigate whether the objective function (1) has a unique minimum or not. Unfortunately, unlike the unweighted case where there is unique global minimum provided the singular values of M are different, (1) might have more local minima. For example, in case of rank–one approximation of $D = \begin{bmatrix} 1 & 1.1 \\ 1 & -1 \end{bmatrix}$ with weight matrix $W = \begin{bmatrix} 1+\alpha & 1 \\ 1 & 1+\alpha \end{bmatrix}$, $\alpha \geq 0$, for large α a non–global local minimum appears, see [18].

III. STOCHASTIC MODELS

The low–rank approximation (1) has the following reformulation to obtain a stochastic problem. Suppose that the data matrix D can be decomposed to the sum of low–rank signal matrix and, in general, full–rank noise matrix in the following way

$$D = UV^\top + E, \tag{3}$$

where $U \in \mathbb{R}^{I\times K}$, $V \in \mathbb{R}^{J\times K}$, and $E = (\varepsilon_{ij})$ is a random matrix of $I \times J$ order. In the sequel, we assume that ε_{ij}'s are mutually independent normally distributed random variables with mean zero and known variances w_{ij}^{-1}, where $w_{ij} > 0$, for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$. The model (3) will be called heteroscedastic low–rank approximation problem. The assumptions of the model imply that the likelihood of the data can be written as an exponential of a quadratic form. Namely, the loglikelihood ℓ can be expressed by a weighted sum of squares in the form

$$\ell(U, V) \propto -Q(U, V), \tag{4}$$

where Q is defined in (1). We therefore see that maximizing likelihood is equivalent, as far as the error variances are known, to minimizing the weighted sum of squares defined in (1) solving a weighted low–rank approximation problem. There are at least two reasons to consider weighted approximation problems allowing the noise variance to be heteroscedastic. On the one hand, we would like to express our differing uncertainty over the signal value of each data entry by the error variance $\sigma_{ij}^2 := w_{ij}^{-1}$. On the other hand, we may have external knowledge derived from experts on the precision or realibility of data entries. Thus, the weight w_{ij} is referred as the precision of the data entry d_{ij} at ith row and jth column. The matrix $W := (w_{ij})$ is called precision matrix.

Another way to derive a model which leads to the heteroscedastic probabilistic model (3) is to consider a homoscedastic low–rank approximation problem under longitudinal study. Let us assume that we have multiple observations d_{ij}^ℓ, $\ell = 1, \ldots, n_{ij}$ ($n_{ij} \in \mathbb{Z}_+$) in each cell (i, j), where $i = 1, \ldots, I$, $j = 1, \ldots, J$, which satisfy the probabilistic model

$$d_{ij}^\ell = \sum_{k=1}^{K} u_{ik} v_{jk} + \varepsilon_{ij}^\ell, \tag{5}$$

where $U = (u_{ik})$ and $V = (v_{jk})$ are parameter matrices of $I \times K$ and $J \times K$ order, respectively, and ε_{ij}^ℓ are mutually independent normally distributed random variables with mean zero and variance σ^2. Then, introducing $\overline{D} = (\overline{d}_{ij})$ with $\overline{d}_{ij} := n_{ij}^{-1} \sum_{\ell=1}^{n_{ij}} d_{ij}^\ell$ and $\overline{E} = (\overline{\varepsilon}_{ij})$ with $\overline{\varepsilon}_{ij} := n_{ij}^{-1} \sum_{\ell=1}^{n_{ij}} \varepsilon_{ij}^\ell$, we have

$$\overline{D} = UV^\top + \overline{E}. \tag{6}$$

This model is a particular case of model (3) since $\mathsf{Var}(\overline{\varepsilon}_{ij}) = \sigma^2/n_{ij}$. The weight matrix for model (6) is given by $W := (n_{ij}/\sigma^2)$.

In certain situations we might consider non–negative integer–valued data matrix which comes from independent Bernoulli experiments possessing similar arrangement to the previous longitudinal study. Denote by $\mathsf{Be}(p)$ the Bernoulli probability distribution with mean $p \in [0, 1]$, i.e., we write $\xi \sim Be(p)$ for a random variable ξ if $\mathsf{P}(\xi = 1) = 1 - \mathsf{P}(\xi = 0) = p$. Assume that the data d_{ij}^ℓ, $\ell = 1, \ldots, n_{ij}$ ($n_{ij} \in \mathbb{Z}_+$), $i = 1, \ldots, I$ and $j = 1, \ldots, J$, satisfy

$$d_{ij}^\ell \sim \mathsf{Be}(g(a_{ij})) \quad \text{with} \quad A = (a_{ij}) = UV^\top,$$

where g is the logistic function defined by $g(a) := (1 + \exp(-a))^{-1}$, $a \in \mathbb{R}$, U and V are parameter matrices of

$I\times K$ and $J\times K$ order, respectively. Let us introduce the data matrix $D=(d_{ij})$, where $d_{ij}:=\sum_{\ell=1}^{n_{ij}} d_{ij}^{\ell}$. Then $D\in\mathbb{Z}_+^{I\times J}$, d_{ij}'s are mutually independent, and they have binomial distribution with parameters n_{ij} and $g(a_{ij})$, $i=1,\dots,I$ and $j=1,\dots,J$. We will refer to this model in the following way

$$D\sim \mathsf{Bi}(N,g(UV^\top)), \tag{7}$$

where $N:=(n_{ij})$, the function g acts on a matrix A entry–wise, i.e., $g(A):=(g(a_{ij}))$, and $\mathsf{Bi}(N,P):=(\mathsf{Bi}(n_{ij},p_{ij}))$ for $N\in\mathbb{Z}_+^{I\times J}$, $P\in[0,1]^{I\times J}$, where the entries are mutually independent. The parameter (U,V) is estimated by the maximum likelihood method. For the loglikelihood we have

$$\ell(A)=\langle D,A\rangle_F-\langle N,\ln(1+\exp A)\rangle_F, \tag{8}$$

where $\exp A:=(\mathrm{e}^{a_{ij}})$. If we define $\widetilde{d}_{ij}^{\ell}:=2d_{ij}^{\ell}-1$ then we obtain that

$$\ell(A)=-\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{\ell=1}^{n_{ij}}\ln\left(1+\exp\left(-\widetilde{d}_{ij}^{\ell}a_{ij}\right)\right),$$

where the sum after the negative sign is the so–called logistic loss. Thus, maximizing the loglikelihood is equivalent to minimizing the logistic loss. The logistic low–rank approximation problem is to maximize the objective function $\ell(A)$ subject to the low–rank constraint $\mathsf{rank}(A)=K$. Like for the heteroscedastic low–rank approximation problem (3) the decomposition $UV^\top$ of A is not unique. However, taking the SVD of this matrix we have unique decomposition and thus parametrization for model (7) provided the singular values of A are distinct. In general, we will refer to the optimization of the objective function $\ell(U\Sigma V^\top)$ subject to $U\in\mathbb{R}^{I\times K}$, $V\in\mathbb{R}^{J\times K}$ are orthogonal matrices and $\Sigma\in\mathbb{R}_+^{K\times K}$ is a diagonal matrix as the K-th order logistic singular valued decomposition (LSVD) of pair (D,N).

In fact, the models (3) and (7) are particular cases of the following general stochastic low–rank approximation problem. Let $f(\cdot|\theta)$, $\theta\in\mathbb{R}$, be a known family of generalized probability density functions with respect to a σ–finite measure μ over a probability space. Let us given a set of mutually independent observations d_{ij}^{ℓ} which have generalized probability density functions $f(\cdot|\theta_{ij})$ for all $\ell=1,\dots,n_{ij}$ ($n_{ij}\in\mathbb{Z}_+$), $i=1,\dots,I$ and $j=1,\dots,J$. We suppose that the parameter matrix $\Theta:=(\theta_{ij})$ satisfies a low–rank condition. Then the stochastic low–rank matrix approximation problem is to estimate Θ based on the data d_{ij}^{ℓ}'s.

IV. ASYMPTOTIC INFERENCE

The partial derivatives of the loglikelihood (4) become, denoting by $\circ$ the entry–wise multiplication,

$$\begin{aligned}\frac{\partial\ell}{\partial U}&=(W\circ(D-UV^\top))V,\\ \frac{\partial\ell}{\partial V}&=(W^\top\circ(D^\top-VU^\top))U.\end{aligned} \tag{9}$$

The partial derivatives should vanish at the maximum likelihood estimator $(\widehat{U},\widehat{V})$ of the model (3), i.e.

$$\nabla\ell:=\left(\frac{\partial\ell}{\partial U},\frac{\partial\ell}{\partial V}\right)=0, \tag{10}$$

which is bilinear in U and V. Due to the nonlinear nature of this estimating equation numerical optimization methods are needed to find its solution. A few of them are described in the next two sections.

The Fisher information in this case will be a linear mapping on $\mathbb{R}^{(I+J)\times K}$. Its matrix representation in the standard basis is given as a block matrix

$$\mathcal{F}(U,V)=\begin{bmatrix}\mathcal{F}_{UU}&\mathcal{F}_{UV}\\ \mathcal{F}_{VU}&\mathcal{F}_{VV}\end{bmatrix},$$

where $\mathcal{F}_{UU}$ is a linear mapping on $\mathbb{R}^{I\times K}$ with the following matrix in the standard basis

$$\mathcal{F}_{UU}(i,k;m,l):=\mathsf{E}\left(\frac{\partial\ell}{\partial u_{ik}}\frac{\partial\ell}{\partial u_{ml}}\right),$$

$i,m=1,\dots,I$, $k,l=1,\dots,K$, and the matrices $\mathcal{F}_{UV}$, $\mathcal{F}_{VU}$ and $\mathcal{F}_{VV}$ are defined similarly. For a pair (A,B), where $A\in\mathbb{R}^{I\times K}$ and $B\in\mathbb{R}^{J\times K}$, the image $\mathcal{F}(A,B)$ of (A,B) is the pair $(\mathcal{F}_{UU}A+\mathcal{F}_{UV}B,\mathcal{F}_{VU}A+\mathcal{F}_{VV}B)$. We derive explicit formula for Fisher information expressing it by U, V and W. Introduce the matrices

$$\begin{aligned}U_{kl}&=\mathsf{diag}\{\langle u_k,u_l\rangle_{w_j},j=1,\dots,J\},\\ V_{kl}&=\mathsf{diag}\{\langle v_k,v_l\rangle_{w^i},i=1,\dots,I\},\end{aligned}$$

for all $k,l=1,\dots,K$. Then we have, for the proof see the Appendix, that

$$\begin{aligned}\mathcal{F}_{UU}&=\mathsf{mat}\{V_{kl},k,l=1,\dots,K\},\\ \mathcal{F}_{VV}&=\mathsf{mat}\{U_{kl},k,l=1,\dots,K\}.\end{aligned} \tag{11}$$

Moreover

$$\begin{aligned}\mathcal{F}_{UV}&=\mathsf{mat}^\top\{W\circ(u_kv_l^\top),k,l=1,\dots,K\},\\ \mathcal{F}_{VU}&=\mathsf{mat}\{W^\top\circ(v_ku_l^\top),k,l=1,\dots,K\},\end{aligned} \tag{12}$$

where $\mathsf{mat}^\top$ denotes the block–transpose. It is well–known that the Fisher information is positive semi–definit. However, since the model is overparametrized it is not positive definit and hence it is not invertable. In the next lemma we characterize the null space of the Fisher information, for the proof see the Appendix.

Lemma 1: Suppose that the matrices U and V are of full rank K and the weight matrix W is strictly positive. Then the null space of the Fisher information $\mathcal{F}(U,V)$ is the K^2–dimensional subspace

$$\ker(\mathcal{F}(U,V))=\left\{(UR,-VR^\top)\,|\,R\in\mathbb{R}^{K\times K}\right\}. \tag{13}$$

Recall that if an estimator $(\widehat{U},\widehat{V})$ is a solution of the likelihood equation (10) then it is called M–estimator. The covariance matrix of an M–estimator $(\widehat{U},\widehat{V})$ is also singular since its distribution is concentrated on a lower dimensional subspace. In order to investigate the asymptotic behaviour of M–estimators of model (3) we have to consider a sequence

of these models with same parameter (U,V) under appropriate sequence of error matrices. Since by Lemma 1 the Fisher information matrix is singular we could not expect that an M–estimator $(\widehat{U},\widehat{V})$ converges weakly to a non–degenerated distribution under any normalization. However, if we consider a proper parametrization, e.g., the parametrization introduced in Section II, then the following asymptotic result holds.

Theorem 1: Suppose that the sequence of data matrices D_n, $n\in\mathbb{N}$, satisfies the model (3) with parameter (U,V) and precision matrices W_n, $n\in\mathbb{N}$, such that $n^{-1}W_n\to W$ as $n\to\infty$, where W is a strictly positive matrix. Moreover, let consider a proper parametrization $\theta\in\mathbb{R}^{(I+J-K)K}$ of the model. If $\widehat{\theta}_n$ is a maximum likelihood estimator of θ based on D_n for all $n\in\mathbb{N}$, then the sequence $\sqrt{n}(\widehat{\theta}_n-\theta)$, $n\in\mathbb{N}$, is asymptotically normal with mean zero and covariance matrix $\mathcal{F}^{-1}(\theta)$, where $\mathcal{F}(\theta)$ is the positive definit Fisher information matrix at θ.

The proof is based on the general results of Section 5.5. in [22]. One can easily check that model (5) satisfies the condition of Theorem 1 provided $n_{ij}/n\to w_{ij}>0$ for all $i=1,\dots,I$ and $j=1,\dots,J$ as $n\to\infty$. We conjecture that similar result holds for the logistic low–rank model (7).

In practice, we are interested in the entries of the approximate matrix M but not the parameter θ itself. The entries of M depend on the parameter θ by a differentiable function hence their asymptotical behaviour can be described by the delta method. More general, if $g:\mathbb{R}^{(I+J-K)K}\to\mathbb{R}^L$ is a differentiable function at θ, where $L\in\mathbb{N}$, then the sequence $\sqrt{n}(g(\widehat{\theta}_n)-g(\theta))$, $n\in\mathbb{N}$, is asymptotically normal with mean zero and covariance matrix $g'(\theta)\mathcal{F}^{-1}(\theta)(g'(\theta))^\top$, see Theorem 3.1 in [22]. It is hard to handle the proper parametrization and to deduce explicit formula for Fisher information matrix. Hence we revert to the original parametrization given by the pair (U,V). One can see that there exists a differentiable function $h:\mathbb{R}^{(I+J)K}\to\mathbb{R}^L$ such that $h(U,V)=g(\theta)$ if $M(\theta)=UV^\top$. Then the asymptotic covariance matrix of $h(\widehat{U}_n,\widehat{V}_n)=g(\widehat{\theta}_n)$, $n\in\mathbb{N}$, can be expressed by the Moore–Penrose inverse of the Fisher information $\mathcal{F}(U,V)$ in the form $h'(U,V)\mathcal{F}^-(U,V)(h'(U,V))^\top$.

V. WEIGHTED DYADIC APPROXIMATION

In this section we consider the rank–one case. Suppose that the data matrix D satisfies the stochastic model

$$D=uv^\top+E, \tag{14}$$

where $u\in\mathbb{R}^I$ and $v\in\mathbb{R}^J$ are parameter vectors, and the entries ε_{ij} of the error matrix E are mutually independent normally distributed random variables with mean zero and precision w_{ij} for all $i=1,\dots,I$ and $j=1,\dots,J$. As we mentioned in Section II this model is not identifiable since αu and $\alpha^{-1}v$ is an equivalent parametrization for all $\alpha\neq 0$. A proper parametrization can be derived in the following way. Without loss of generality we may suppose that $u_1v_1\neq 0$. Let $o:=(u_1v_1)/|u_1v_1|$, $\alpha:=(|u_1v_1|)^{1/2}$, $\tilde{u}:=o\alpha^{-1}v_1(u_2,\dots,u_r)^\top$ and $\tilde{v}:=o\alpha^{-1}u_1(v_2,\dots,v_r)^\top$. Then we have

$$uv^\top=o\begin{bmatrix}\alpha\\ \tilde{u}\end{bmatrix}[\alpha\,\tilde{v}]$$

and $\vartheta:=(\alpha,\tilde{u}^\top,\tilde{v}^\top)^\top$ is a $I+J-1$ dimensional parametrization. Clearly, the two congruence classes in the set of dyads are given according to the value of o, which is $+1$ or -1. However, in this proper parametrization the data matrix depends quadratically in α that is hard to handle. Hence we will use the original bilinear parametrization (u,v) inspite of its redundancy.

The loglikelihood of the model (14) is given by $\ell(u,v)\propto -Q(u,v)$, where Q is defined in (1). For a parameter (u,v) let us define the estimated error matrix by

$$\widehat{E}(u,v):=D-uv^\top. \tag{15}$$

By (9), for the partial derivatives of ℓ we have

$$\frac{\partial\ell}{\partial u}=(W\circ\widehat{E}(u,v))v=(W\circ D)v-R(v)u,$$
$$\frac{\partial\ell}{\partial v}=(W\circ\widehat{E}(u,v))^\top u=(W\circ D)^\top u-C(u)v,$$

where $C(u)=\mathsf{diag}\{\langle u,u\rangle_{w_j},j=1,\dots,J\}$ and $R(v)=\mathsf{diag}\{\langle v,v\rangle_{w^i},i=1,\dots,I\}$. The stationary points of ℓ can be derived by solution of the equation $\nabla\ell=0$, where the gradient is defined by $\nabla\ell:=(\frac{\partial\ell}{\partial u},\frac{\partial\ell}{\partial v})$. We have

$$R(v)u=(W\circ D)v,$$
$$C(u)v=(W\circ D)^\top u.$$

One of the possible way to solve these nonlinear equations is the criss–cross regression or the method of nonlinear iterative partial least squares (NIPALS) suggested in [23]. We iterate the following steps: fixing v we solve the first linear equation onto u, then we do same for v in the second equation by fixing u.

The Fisher information in this case can be expressed as a block matrix of $(I+J)\times(I+J)$ order

$$\mathcal{F}(u,v)=\begin{bmatrix}R(v) & W\circ(uv^\top)\\ W^\top\circ(vu^\top) & C(u)\end{bmatrix}.$$

By Lemma 1 the null space of this matrix is the one–dimensional subspace generated by the vector $(u,-v)\in\mathbb{R}^{I+J}$. The range of the Fisher information matrix is the $I+J-1$–dimensional subspace $\mathcal{R}(u,v):=\{(x,y)|x\in\mathbb{R}^I,y\in\mathbb{R}^J:u^\top x=v^\top y\}$. The second order partial derivatives of ℓ are given by $\frac{\partial^2\ell}{\partial u\partial u^\top}=-R(v)$, $\frac{\partial^2\ell}{\partial v\partial v^\top}=-C(u)$, and $\frac{\partial^2\ell}{\partial u\partial v^\top}=W\circ D-2W\circ(uv^\top)$. Then the Hessian of the loglikelihood ℓ can be written as

$$\mathcal{H}(u,v)=-\mathcal{F}(u,v)+\mathcal{E}(u,v), \tag{16}$$

where

$$\mathcal{E}(u,v):=\begin{bmatrix}0 & W\circ\widehat{E}(u,v)\\ W^\top\circ\widehat{E}^\top(u,v) & 0\end{bmatrix}.$$

Considering a Taylor series expansion around (u, v), we have for all $x \in \mathbb{R}^I$ and $y \in \mathbb{R}^J$

$$\ell(u+x,v+y) = \ell(u,v) + \nabla\ell^\top(u,v)\begin{bmatrix}x\\y\end{bmatrix} + \frac{1}{2}[x^\top\, y^\top]\mathcal{H}(u+\alpha x, v+\alpha y)\begin{bmatrix}x\\y\end{bmatrix}$$

where $\alpha \in (0,1)$. We suggest a trust–region method, see Chapter 4 in [16], for numerical optimization of the loglikelihood. Trust–region methods define a region around the current iterate within which the model is an adequate approximation of the objective function. We may assume that the second part of the Hessian in (16) is negligible and define the quadratic model function m for approximating ℓ around (u, v) as

$$m(x,y) = \ell(u,v) + \nabla\ell^\top(u,v)\begin{bmatrix}x\\y\end{bmatrix} - \frac{1}{2}[x^\top\, y^\top]\mathcal{F}(u,v)\begin{bmatrix}x\\y\end{bmatrix}.$$

Since

$$\nabla\ell^\top(u,v)\begin{bmatrix}u\\-v\end{bmatrix} = [u^\top v^\top]\mathcal{E}(u,v)\begin{bmatrix}u\\-v\end{bmatrix} = 0$$

and the vector $(u, -v)$ belongs to the null space of $\mathcal{F}(u, v)$ we may restrict the increment vector at an iteration onto the subspace $\mathcal{R}(u, v)$. Thus, the model function m is a good approximation of ℓ around (u, v) in the trust–region $\mathcal{R}(u, v)$. The minimum of m is given as the solution of the linear equation

$$\mathcal{F}(u,v)\begin{bmatrix}x\\y\end{bmatrix} = \nabla\ell(u,v) \tag{17}$$

whose matrix is singular by Lemma 1. We can solve this equation by applying Schur complements. Define the Schur complements of the Fisher information in the following way

$$S_R(u,v) := R(v) - (W \circ (uv^\top))C^{-1}(u)(W^\top \circ (vu^\top)),$$
$$S_C(u,v) := C(u) - (W^\top \circ (vu^\top))R^{-1}(v)(W \circ (uv^\top)).$$

We have the following lemma for Schur complements.

Lemma 2: The Schur complements $S_R(u, v)$ and $S_C(u, v)$ are positive semidefinite and their null space are generated by u and v, respectively.

By the block structure of Fisher information and gradient (17) consists of two linear equations for the increments x and y. After expressing y by the second equation and substituting into the first equation we obtain that x is given by the solution of the linear equation

$$S_R(u,v)x = (W \circ \widehat{E}(u,v))v - (W \circ (uv^\top))\times C^{-1}(u)(W^\top \circ \widehat{E}^\top(u,v))u. \tag{18}$$

Then y can be expressed as the following function of x

$$y = C^{-1}(u)\left((W^\top \circ \widehat{E}^\top(u,v))u - (W^\top \circ (vu^\top))x\right).$$

Since, by Lemma 2, the equation (18) is singular we have to eliminate this singularity by Gram–Schmidt orthonormalization. By the Gram–Schmidt process we may find an ortogonal matrix $O_u \in \mathbb{R}^{I\times I}$ whose first column is the unit vector $u/\|u\|$. Let us denote the matrix $(0, I_{I-1})$ of order $(I-1)\times I$ by J and introduce the vector $\widetilde{x} := JO_u x$ and the matrix $\widetilde{S}_R(u,v) := JO_u S_R(u,v)O_u^\top J^\top$. Then $\widetilde{x}$ fulfills the equation

$$\widetilde{S}_R(u,v)\widetilde{x} = JO_u\big(W \circ \widehat{E}(u,v))v - (W \circ (uv^\top))\times C^{-1}(u)(W^\top \circ \widehat{E}^\top(u,v))u\big). \tag{19}$$

Since $\widetilde{S}_R(u, v)$ is a symmetric positive definite matrix this equation has a unique solution, and we can solve it by the Cholesky decomposition of $\widetilde{S}_R(u, v)$. Then x is given by $x = O_u^\top J^\top \widetilde{x}$. From the point of view of the computational cost this order for solving equation (17) is prefered when $I \le J$ else we propose to solve (17) first in y by using Schur complement $S_C(u, v)$.

Summarizing, the weighted rank–one approximation algorithm by Fisher scoring is provided below.

ALGORITHM A: Weighted rank–one approximation
INPUT: Data matrix D, weight matrix W.
1. Choose an initial setting for u and v.
REPEAT
2. Evaluate the error matrix $\widehat{E}(u, v)$ by (15).
3. Evaluate the diagonal matrices $C(u)$ and $R(v)$.
4. Evaluate the Schur complement $S_R(u, v)$.
5. Orthogonalize (u, I_I) to obtain O_u.
6. Solve (19) by Cholesky decomposition of $\widetilde{S}_R(u, v)$.
7. Compute x and y.
8. Update: let $u = u + x$ and $v = v + y$.
UNTIL the convergence criterion is satisfied.
OUTPUT: Rank–one matrix $uv^\top$.

VI. LOW–RANK APPROXIMATION ALGORITHMS

In this section we consider the general low–rank case. It is still possible to obtain satisfying low–rank approximation by means of stepwise dyadic residual fitting or successive dyadic fits, see [10], e.g., using our Algorithm A recursively or several times. However, we propose a new algorithm which takes into account the matricial nature of parameters U and V.

The algorithm is based on the quadratic model function M defined by

$$M(X,Y) = \ell(U,V) + \left\langle \nabla\ell(U,V), \begin{bmatrix}X\\Y\end{bmatrix}\right\rangle_F - \left\langle \frac{1}{2}\begin{bmatrix}X\\Y\end{bmatrix}, \mathcal{F}(U,V)\begin{bmatrix}X\\Y\end{bmatrix}\right\rangle_F,$$

which is a satisfying approximation of ℓ around (U, V) provided the increment $(X, Y) \perp \ker(\mathcal{F}(U, V))$. The minimum of M is given as the solution of the linear equation

$$\mathcal{F}(U,V)\begin{bmatrix}X\\Y\end{bmatrix} = \begin{bmatrix}(W \circ \widehat{E}(U,V))V\\(W \circ \widehat{E}(U,V))^\top U\end{bmatrix}, \tag{20}$$

where the error matrix is defined by $\widehat{E}(U, V) := D - UV^\top$. One can solve this linear equation by taking the Schur complement $\mathcal{S}(U, V) := \mathcal{F}_{UU} - \mathcal{F}_{UV}\mathcal{F}_{VV}^{-1}\mathcal{F}_{VU}$ and introducing

its non–singular transform $\widetilde{S}$ similarly to the preceding section. Thus, the weighted low–rank approximation algorithm based on the Fisher scoring is the following.

ALGORITHM B: Weighted low–rank approximation with
INPUT: Data matrix D, weight matrix W, rank K.
1. Choose an initial setting for U and V.
REPEAT
2. Evaluate the error matrix $\widehat{E}(U,V)$.
3. Evaluate the diagonal block matrices $\mathcal{F}_{UU}$ and $\mathcal{F}_{VV}$.
4. Evaluate the Schur complement $\mathcal{S}(U,V)$.
5. Orthogonalize (X, I_{IK}) in the Frobenius norm.
6. Solve (20) by Cholesky decomposition of $\widetilde{\mathcal{S}}(u,v)$.
7. Compute X and Y.
8. Update: let $U = U + X$ and $V = V + Y$.
UNTIL the convergence criterion is satisfied.
OUTPUT: Rank–K matrix $UV^\top$.

Finally, we move on to the logistic low–rank approximation problem. Consider again the Taylor expansion of ℓ defined by (8) up to second order. We have that the quadratic model function

$$Q(A) := \frac{1}{2}\left\|A - (UV^\top + W^{\circ -1}(D - N \circ P))\right\|_W^2$$

is a good approximation of the negative loglikelihood around $UV^\top$, where $P = P(U,V) := g(UV^\top)$, the weight matrix $W := N \circ P \circ (E - P)$ with $E = (1_{ij})$, and $W^{\circ -1} := (w_{ij}^{-1})$. We may formulate an EM algorithm for solving the logistic low–rank approximation problem as follows.

ALGORITHM C: Logistic low–rank approximation
INPUT: Data matrices D, N, rank K.
1. Choose an initial setting for U and V.
REPEAT
E step. Compute $P(U,V)$ and $W(U,V)$.
M step. Solve the weighted low–rank problem $Q(UV^\top)$.
UNTIL the convergence criterion is satisfied.
OUTPUT: Rank–K matrix $UV^\top$.

At the M step the Algorithm B is used. If we take the SVD of the output $UV^\top$ then we obtain the logistic SVD of the pair (D, N). Thus, we may refer to Algorithm C as iteratively reweighted singular valued decomposition (IRSVD).

VII. APPLICATION TO HUNGARIAN MORTALITY DATA

For mortality tables n_{ij} and d_{ij} denote the total number of individuals at the beginning of the year and the number of deaths in the year, respectively, at age i in year j. (I is the maximal age and J denotes the number of years.) For Hungarian mortality data the observed period is 1949-2008, i.e. $I = 60$, and $J = 101$ supplied by the Hungarian Central Statistical Office. Having looked at the singular values given by the standard SVD it turned out that the best candidate for the low rank is $K = 2$. Figure 1 plots the Hungarian mortality data for women (left) and shows the mortality estimation (right) derived by logistic rank–2 approximation.

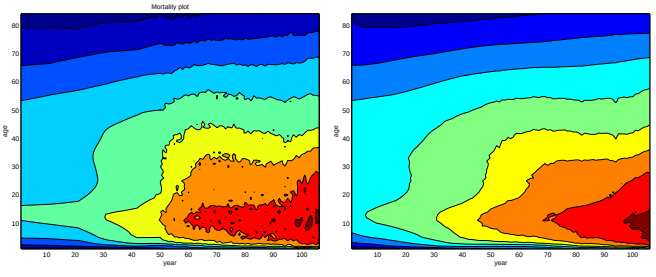

Fig. 1. Hungarian mortality data (left), estimated (right)

VIII. APPENDIX

Proof of block–matrix representation (2). The matrix M has SVD in the form $M = U\Sigma V^\top$ with orthogonal matrices $U \in \mathbb{R}^{I\times I}$, $V \in \mathbb{R}^{J\times J}$ and diagonal matrix $\Sigma \in \mathbb{R}^{I\times J}$, see Theorem 4.1 in [21]. Since $\mathsf{rank}(M) = K$ we have $\mathsf{rank}(\Sigma) = K$, i.e., $\Sigma = \mathsf{diag}\{\sigma_1, \dots, \sigma_K, 0, \dots, 0\}$ with $\sigma_i > 0$ for all $i = 1, \dots, K$. Consider the block–matrix representation of U and V:

$$U = \begin{bmatrix} U_1 & U_2 \\ U_3 & U_4 \end{bmatrix}, \qquad V = \begin{bmatrix} V_1 & V_2 \\ V_3 & V_4 \end{bmatrix},$$

where $U_1, V_1 \in \mathbb{R}^{K\times K}$. Then, by the orthogonality of U and V, we have $U_1^\top U_1 + U_2^\top U_2 = I_K$ and $V_1^\top V_1 + V_2^\top V_2 = I_K$, and (2) follows by matrix multiplication.

Proof of formula (11). By (3) and (9) we have $\frac{\partial \ell}{\partial U} = (W \circ E)V$, i.e., $\frac{\partial \ell}{\partial u_{ik}} = \sum_{j=1}^J w_{ij}\varepsilon_{ij}v_{jk}$. Hence

$$\begin{aligned}\mathcal{F}_{UU}(i,k;m,l) &= \mathsf{E}\left(\sum_{j=1}^J w_{ij}\varepsilon_{ij}v_{jk}\sum_{n=1}^J w_{mn}\varepsilon_{mn}v_{nl}\right)\\ &= \sum_{j,n}^J w_{ij}w_{mn}v_{jk}v_{nl}\mathsf{E}(\varepsilon_{ij}\varepsilon_{mn})\\ &= \delta_{im}\sum_{j=1}^J w_{ij}v_{jk}v_{jl} = \delta_{im}\langle v_k, v_l\rangle_{w^i}\end{aligned}$$

for all $i, m = 1, \dots, I$, $k, l = 1, \dots, K$. The proof is same for $\mathcal{F}_{VV}$. To prove (12) we note that

$$\begin{aligned}\mathcal{F}_{UV}(i,k;n,l) &= \mathsf{E}\left(\sum_{j=1}^J w_{ij}\varepsilon_{ij}v_{jk}\sum_{m=1}^I w_{mn}\varepsilon_{mn}u_{ml}\right)\\ &= \sum_{m=1}^I\sum_{j=1}^J w_{ij}w_{mn}v_{jk}u_{ml}\mathsf{E}(\varepsilon_{ij}\varepsilon_{mn}) = w_{in}u_{il}v_{nk}.\end{aligned}$$

Proof of Lemma 1. For any pair (A, B), where $A \in \mathbb{R}^{I\times K}$ and $B \in \mathbb{R}^{J\times K}$, we have

$$\begin{aligned}\langle (A,B), \mathcal{F}(A,B)\rangle_F =& \langle A, \mathcal{F}_{UU}A\rangle_F + 2\langle A, \mathcal{F}_{UV}B\rangle_F\\ &+ \langle B, \mathcal{F}_{VV}B\rangle_F.\end{aligned}$$

By definition of the Frobenius norm we have

$$\begin{aligned}\langle A, \mathcal{F}_{UU}A\rangle_F &= \sum_{i,m=1}^{I}\sum_{k,l=1}^{K}\mathcal{F}_{UU}(i,k;m,l)a_{ik}a_{ml}\\ &= \sum_{i,m=1}^{I}\sum_{k,l=1}^{K}\delta_{im}\sum_{j=1}^{J}w_{ij}v_{jk}v_{jl}a_{ik}a_{ml}\\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}w_{ij}\left(\sum_{k=1}^{K}v_{jk}a_{ik}\right)^2\\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}w_{ij}\langle v^j, a^i\rangle.\end{aligned}$$

Similarly, we have

$$\begin{aligned}\langle B, \mathcal{F}_{VV}B\rangle_F &= \sum_{i=1}^{I}\sum_{j=1}^{J}w_{ij}\langle u^i, b^j\rangle,\\ \langle A, \mathcal{F}_{UV}B\rangle_F &= \sum_{i=1}^{I}\sum_{j=1}^{J}w_{ij}\langle u^i, b^j\rangle\langle v^j, a^i\rangle.\end{aligned}$$

Thus

$$\langle (A,B), \mathcal{F}(A,B)\rangle_F = \sum_{i=1}^{I}\sum_{j=1}^{J}w_{ij}\left(\langle u^i, b^j\rangle + \langle v^j, a^i\rangle\right)^2.$$

Since W is strictly positive the right hand side equals to zero if and only if $\langle u^i, b^j\rangle + \langle v^j, a^i\rangle = 0$ for all $i = 1,\dots,I$ and $j = 1,\dots,J$, i.e.,

$$UB^\top + AV^\top = 0. \tag{21}$$

It is easy to see that elements of the subspace (13) fulfill the above equation. It remains to show that if a pair (A,B) satisfies (21) then there exists $R \in \mathbb{R}^{K\times K}$ such that $A = UR$ and $B = -VR^\top$. Taking the SVD of U, since $\mathsf{rank}(U) = K$ there exist matrices $S \in \mathbb{R}^{I\times K}$ and $T, \Sigma \in \mathbb{R}^{K\times K}$ such that $S^\top S = I_K$, $TT^\top = T^\top T = I_K$ and $\Sigma = \mathsf{diag}\{\sigma_1,\dots,\sigma_K\}$ with $\sigma_k > 0$ for all $k = 1,\dots,K$ such that $U = S\Sigma T^\top$. By (21) we have $B = -VA^\top S\Sigma^{-1}T^\top$, i.e., the matrix B can be expressed in the form $-VR^\top$, where $R := T\Sigma^{-1}S^\top A$. Finally, suppose that $B = -VR^\top$ with a matrix $R \in \mathbb{R}^{K\times K}$. Then, by (21), $(A - UR)V^\top = 0$ which implies $A = UR$ since $\mathsf{rank}(V) = K$. This completes the proof.

Proof of Lemma 2. We prove the lemma for the Schur complement $S_R(u,v)$. For all $u, x \in \mathbb{R}^I$ the Cauchy–Schwarz inequality states

$$\left(\sum_{i=1}^{I}w_{ij}u_ix_i\right)^2 \le \sum_{i=1}^{I}w_{ij}u_i^2\sum_{i=1}^{I}w_{ij}x_i^2.$$

Thus, we have

$$x^\top R(v)x = \sum_{i=1}^{I}\sum_{j=1}^{J}w_{ij}x_i^2v_j^2 \ge \sum_{j=1}^{J}v_j^2\frac{\left(\sum_{i=1}^{I}w_{ij}u_ix_i\right)^2}{\sum_{i=1}^{I}w_{ij}u_i^2},$$

where the right hand side is equal to $x^\top(W \circ (uv^\top))C^{-1}(u)(W^\top \circ (vu^\top))x$. Hence $x^\top S_R(u,v)x \ge 0$ for all $x \in \mathbb{R}^I$ and $x^\top S_R(u,v)x = 0$ if and only if there exists $\alpha \in \mathbb{R}$ such that $x = \alpha u$. Thus, $\ker(S_R(u,v)) = \{\alpha u : \alpha \in \mathbb{R}\}$.

References

[1] H.C. Andrews and C.L. Patterson, Outer product expansions and their uses in digital image processing, *Am. Math. Mon.*, vol. 82, 1975, pp 1–13.

[2] S. Baran, J. Gáll, M. Ispány, and G. Pap, Forecasting Hungarian mortality rates using the Lee-Carter method, *Acta Oeconomica*, vol. 57, no. 1, 2007, pp 21–34.

[3] J.P. Benzecri, *Correspondence Analysis Handbook*, Marcel Dekker, New York; 1992.

[4] A. Björck, *Numerical Methods for Least Squares Problems*, SIAM; 1996.

[5] M.W. Berry, M. Browne, A.N. Langville, V.P. Pauca, and R.J. Plemmons, Algorithms and applications for approximate nonnegative matrix factorization, *Comp. Stat. Data Anal.*, vol. 52, 2007, pp 155–173.

[6] J.M. Chambers, W.S. Cleveland, B. Kleiner, P.A. Tukey, *Graphical Methods for Data Analysis*, Wadsworth, Belmont, California; 1983.

[7] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay, Clustering large graphs via the singular valued decomposition, *Machine Learning*, vol. 56, 2004, pp. 9–33.

[8] C. Eckart and G. Young, The approximation of one matrix by another of lower rank, *Psychometrika*, vol. 1, 1936, pp 211–218.

[9] S.E. Fienberg, P. Hersh, A. Rinaldo, and Yi Zhou, "Maximum likelihood estimation in latent class models for contingency table data", *in Algebraic and Geometric Methods in Probability and Statistics*, P. Gibilisco, E. Riccomagno, M.P. Rogantin (eds.), Cambridge University Press, 2009, pp. 27–62.

[10] K.R. Gabriel and S. Zamir, Lower rank approximation of matrices by least squares with any choice of weights, *Technometrics*, vol. 21, no. 4, 1979, pp 489–498.

[11] R.A. Horn and Ch.R. Johnson, *Matrix Analysis*, Cambridge University Press; 1990.

[12] R.D. Lee and L.R. Carter, Modeling and forecasting the time series of U.S. mortality, *J. Amer. Stat. Assoc.*, vol. 87, no. 419, 1992, pp 659–671.

[13] W.S. Lu, S.C. Pei, and P.H. Wang, Weighted low–rank approximation of general complex matrices and its application in the design of 2–D digital filters, *IEEE Trans. on Circuits and Systems*, vol. 44, 1997, pp. 650–655.

[14] D. Lee and H.S. Seung, Learning the parts of objects by non–negative matrix factorization, *Nature*, vol. 401, 1999, pp. 788–791.

[15] J.M. Lee and D.A. Weinberg, A note on canonical forms for matrix congruence, *Lin. Alg. Appl.*, vol. 249, 1996, pp 207–215.

[16] J. Nocedal, S.J. Wright, *Numerical Optimization*, Springer; 1999.

[17] S.S. Schiffman, M.L. Reynolds, and F.W. Young, *Introduction to Multidimensional Scaling*, Academic Press, New York; 1981.

[18] N. Srebro and T. Jaakkola, "Weighted low–rank approximation", *in Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003)*, T. Fawcett and N. Mishra (eds), Washington DC, USA, AAAI Press, 2003, pp 720–727.

[19] X. Su and T.M. Khoshgoftaar, A survey of collaborative filtering techniques, *Advances in Artifical Intelligence*, vol. 2009, 2009, pp. 1–19.

[20] Y. Tian, T. Tan, Y Wang, and Y Fang, Do singular values contain adequate information for face recognition, *Pattern Recognition*, vol. 36, 2003, 649–655.

[21] L.N. Trefethen and D. Bau, III, *Numerical Linear Algebra*, SIAM; 1997.

[22] A.W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press; 1998.

[23] H. Wold and E. Lyttkens, Nonlinear iterative partial least squares (NIPALS) estimation procedures, *Bull. Inter. Statist. Inst.*, vol. 43, 1969, 29–51.