ON THE LOW-RANK APPROXIMATION OF DATA ON THE UNIT SPHERE

M. CHU*, N. DEL BUONO†, L. LOPEZ‡, AND T. POLITI§

Abstract. In various applications, data in multi-dimensional space are normalized to unit length. This paper considers the problem of best fitting given points on the m-dimensional unit sphere S^{m-1} by k-dimensional great circles with k much less than m. The task is cast as an algebraically constrained low-rank matrix approximation problem. Using the fidelity of the low-rank approximation to the original data as the cost function, this paper offers an analytic expression of the projected gradient which, on one hand, furnishes the first order optimality condition and, on the other hand, can be used as a numerical means for solving this problem.

Key words. standardized data, linear model, factor analysis, low-rank approximation, latent semantic indexing, projected gradient

1. Introduction. Given n points on the unit sphere $\mathcal{S}^{m-1}$ in $\mathbb{R}^m$, we concern ourselves in this paper with the question of best fitting these points by a "great circle" which is understood to mean the intersection of $\mathcal{S}^{m-1}$ with a certain k-dimensional subspace of $\mathbb{R}^m$. Because the approximation is by points restricted to $\mathcal{S}^{m-1}$, we are not dealing with an ordinary data fitting problem. Denoting the given points as rows of a data matrix A, the task of best fitting can be recast as a constrained low-rank matrix approximation problem as follows:

Given a matrix $A \in \mathbb{R}^{n\times m}$ whose rows are of unit length, find an approximation matrix $Z \in \mathbb{R}^{n\times m}$ of A whose rows are also of unit length, but $\operatorname{rank}(Z) = k$ *where* $k < \min\{m, n\}$.

Structured low-rank approximation problems arise in many important applications [3, 4, 13, 24, 33]. Some basic theory and computation schemes, including a generic alternating projection algorithm and a robust equality constrained optimization formulation, can be found in a recent article [9] by the first author and the references contained therein. The structure in our current problem assumes the form that each row of Z has unit length. A similar setting appears in an application to designing structured tight frames [34] where the alternating projection method is used as an effective tool. When the row number n is much larger than the rank k, however, reformulations of this problem as the two schemes proposed in [9] would involve significant amount of redundancies in the constraints which, in turn, might cause computational difficulties. The approximation obtained by a linear combination of concept vectors generated by the spherical k-means method [14], on the other hand, lacks the normalization and is limited only to the subspace spanned by the concept vectors. In this paper we propose using the projected gradient approach as an alternative means to tackle the problem. Three types of measurement of best fitting will be used and their results will be compared.

The need of normalizing data before approximations can be performed emerges from many applications. See, for example, the book [12] about pattern classification, the monograph [21] and the vast literature on principal component analysis, the required goal in positive matrix factorization [26], and the spherical k-means method for concept mining in [14]. It might be illuminating to brief readers by some simple examples of why such a normalized low-rank ap-

*Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205. (chu@math.ncsu.edu)
This research was supported in part by the NSF under grants DMS-0073056 and CCR-0204157, and in part by the INDAM under the Visiting Professorship Program when the author visited the Department of Mathematics at the University of Bari.
†Dipartimento di Matematica, Università degli Studi di Bari, Via E. Orabona 4, I-70125 Bari, Italy. (delbuono@dm.uniba.it)
‡Dipartimento di Matematica, Università degli Studi di Bari, Via E. Orabona 4, I-70125 Bari, Italy. (lopezl@dm.uniba.it)
§Dipartimento di Matematica, Politecnico di Bari, Via Amendola 126/B, 70126 Bari, Italy. (politi@poliba.it)

proximation problem is of importance in practice. To set the stage, let $Y = [y_{ij}] \in \mathbb{R}^{n\times\ell}$ denote the matrix of "observed" data. The entry y_{ij} represents, in a broad sense, the *standard score* obtained by entity j on variable i. By a standard score we mean that a raw score per variable has been normalized to have mean 0 and standard deviation 1. After this normalization, the matrix

$$R := \frac{1}{\ell} YY^\top, \tag{1.1}$$

represents the correlation matrix of all n variables. Note that $r_{ii} = 1$ and $|r_{ij}| \le 1$ for all $i, j = 1, \dots n$. In a linear model, it is assumed that the score y_{ij} is a linearly weighted score by entity j based on several factors. We shall temporarily assume that there are m factors, but it is precisely the point that the factors are to be retrieved in the mining process. A linear model, therefore, assumes the relationship

$$Y = AF, \tag{1.2}$$

where $A = [a_{ik}] \in \mathbb{R}^{n\times m}$ is a *loading* matrix with a_{ik} denoting the *influence of factor k on variable i*, and $F = [f_{kj}] \in \mathbb{R}^{m\times\ell}$ is a *scoring* matrix with f_{kj} denoting the *response of entity j to factor k*. Depending on the application, often only a portion of the matrices in equation (1.2) is known and the goal is to retrieve the remaining information. The linear model (1.2) can have different interpretations. We sketch two scenarios below.

Scenario 1. Assume that each of the ℓ columns of the observed matrix Y represents the transcript of a college student (an entity) at the end of freshman year on n fixed subjects (the variables). It is generally believed that a college freshman's academic performance depends on a number of preexistent factors. Upon entering the college, each student could be asked to fill out a questionnaire inquiring his or her backgrounds (scores) with respect to these factors. These responses are placed in the corresponding column of the scoring matrix F. What is not clear to the administrators in the first place is how to choose the factors to compose the questionnaire or how each of the chosen factors would be weighted (the loadings) to reflect their effect on each particular subject. In practice, we usually do not have a priori knowledge about the number and character of underlying factors in A. Sometimes we do not even know the factor scores in F. Only the data matrix Y is observable. Explaining the complex phenomena observed in Y with the help of a *minimal number* of factors extracted from the data matrix is the primary and most important goal of factor analysis.

To collectively gather students' academic records from wide variations (of grading rules and difficulties of subjects) for the purpose of analyzing some influential factors in A, it is essential that all these scores should be evaluated on the same basis. Otherwise, the factors cannot be correctly revealed. Statistically, the same basis of scores means the standardization (of rows) of Y. For our purpose, it is reasonable to assume that all sets of factors being considered are uncorrelated with each other. (See [21] for the case of stochastic independence.) As such, rows of F should be orthogonal to each other. We should further assume that the scores in F for each factor are normalized. Otherwise, one may complain that the factors being considered are not neutral to begin with and the analysis would be biased and misinformed. Under these assumptions, we have

$$\frac{1}{\ell} FF^\top = I_m, \tag{1.3}$$

where I_m stands for the identity matrix in $\mathbb{R}^{m\times m}$. It follows that the correlation matrix R can be expressed directly in terms of the loading matrix A alone, i.e.,

$$R = AA^\top. \tag{1.4}$$

Factor extraction now becomes a problem of decomposing the correlation matrix R into the product $AA^\top$ using as few factors as possible [18]. *Due to the normalization we point out that the rows of A always are of unit length.*

The kth column of A may be interpreted as correlations of the data variables with that particular kth factor. Those data variables with high factor loadings are considered to be more influenced by the factor in some sense and those with zero or near-zero loadings are treated as being unlike the factor. The quality of this likelihood, indicating the *significance* of the corresponding factor, is measured by the norm of the kth column of A. One basic idea in factor analysis is to rewrite the loadings of variables over some newly selected factors so as to manifest more clearly the correlation between variables and factors [17, 21]. Those factors with smaller significance are likely to be dropped to reduce the number of effectual factors. *This notion of factor reduction amounts to a low-rank approximation of A.*

Scenario 2. Assume now that the textual documents are collected in an *indexing matrix* $H = [h_{ik}]$ in $\mathbb{R}^{n\times m}$. Each document is represented by one row in H. The entry h_{ik} represents the *weight* of one particular *term* k in document i whereas each term could be defined by just one single word or a string of phrases. A natural choice of the weight h_{ik} is by counting the number of times that the term k occurs in document i. Many more elaborate weighting schemes that yield better performance have been devised in the literature. See, for example, lists in [23, 28] and comparisons in [31]. One commonly used term-weighting scheme to enhance discrimination between various documents and to enhance retrieval effectiveness is to define

$$h_{ik} = t_{ik} g_k n_i, \tag{1.5}$$

where t_{ik} captures the relative importance of term k in document i, g_k weights the overall importance of term k in the entire set of documents, and

$$n_i = \left(\sum_{k=1}^{m} t_{ik} g_k \right)^{-1/2} \tag{1.6}$$

is the scaling factor for normalization. The normalization by (1.6) is necessary because, otherwise, one could artificially inflate the prominence of document i by padding it with repeated pages or volumes. *After the normalization, one sees that the rows of H are of unit length.*

One usage of the indexing matrix is for information retrieval, particularly in the context of latent semantic indexing (LSI) application [2, 16]. Each query is represented as a column vector $\mathbf{q}_j^\top = [q_{1j}, \ldots, q_{mj}]$ in $\mathbb{R}^m$ where q_{kj} represents the weight of term k in the query j. Similar to (1.5), more elaborate schemes can also be employed to weight terms in a query. To measure how the query $\mathbf{q}_j$ matches the documents, we calculate the column vector

$$\mathbf{c}_j = H\mathbf{q}_j \tag{1.7}$$

and rank the relevance of documents to $\mathbf{q_j}$ according to the *scores* in $\mathbf{c}_j$.

To put the notation in the context of linear model (1.2), we observe the follow analogies between the two examples:

$$\begin{aligned}
\text{indexing matrix } H &\longleftrightarrow \text{loading matrix } A \\
\text{document } i &\longleftrightarrow \text{variable } i \\
\text{term } k &\longleftrightarrow \text{factor } k \\
\text{one query } \mathbf{q}_j &\longleftrightarrow \text{one column in scoring matrix } F \\
\text{weight } h_{ik} \text{ of term } k \text{ in document } i &\longleftrightarrow \text{loading } a_{ik} \text{ of factor } k \text{ on variable } i \\
\text{weights } q_{kj} \text{ of term } k \text{ in query } \mathbf{q}_j &\longleftrightarrow \text{response } f_{kj} \text{ of entity } j \text{ to factor } k \\
\text{rank } \mathbf{c}_{ij} \text{ of document } i \text{ in query } \mathbf{q}_j &\longleftrightarrow \text{score } y_{ij} \text{ of variable } i \text{ in entity } j
\end{aligned}$$

It is important to note the different emphasis in the two examples. In Scenario 1 the data matrix Y is given and the factors in A are to be retrieved while in Scenario 2 the terms in the indexing matrix H are predetermined and the scores $\mathbf{c}_j$ are to be calculated.

The computation involved in the LSI application is more on the vector-matrix multiplication (1.7) than on factor retrieval, provided that H is a "reasonable" representation of the relationship between documents and terms. In practice, however, the matrix H is never exact nor perfect. How to represent the indexing matrix and the queries in a more compact form so as to facilitate the computation of the scores therefore become a major challenge in the field [14, 27]. One idea, with the backing of both statistical analysis [7] and empirical testing [2], is to represent H by its low-rank approximation. A constrained low-rank approximation problem therefore arises. We shall point out in our numerical examples that the fidelity to H by merely normalizing a truncated singular value decomposition is not as high as what our method can achieve.

One final point connecting Scenario 1 to Scenario 2 is worth mentioning. It is plausible that once we have accumulated enough experiences $C = [\mathbf{c}_1, \ldots, \mathbf{c}_\ell]$ from queries $\mathbf{q}_1, \ldots, \mathbf{q}_\ell$, we should be able to employ whatever factor retrieval techniques developed for Scenario 1 to reconstruct terms (factors) for Scenario 2. That is, after the standardization of the scores, the information gathered in C may be used as a feedback to the process of selecting newer and more suitable H. Recall from (1.4) that no reference to the queries is needed. This *learning process* with feedback from C can be iterated to reduce and improve the indexing matrix H.

2. Low Rank Approximation. The linear model (1.2) can have a much broader interpretation than the two applications we have demonstrated. Its interpretation in Scenario 1 for factor analysis manifests the critical role that the loading matrix A will play. Our low-rank approximation problem means the need to retrieve a few principal factors to represent the original A. Likewise, the interpretation in Scenario 2 for the information retrieval suggests that the indexing matrix H is playing an equally important role as the loading matrix A. Our low-rank approximation problem in the context of LSI means the need to derive an approximation of H to allow fast computation. In this section, we introduce a notion that can be employed to deal with this algebraically constrained low-rank approximation problem.

The quality of the approximation by a lower rank matrix Z to A can be measured in different ways. In the following, we shall describe three kinds of measurements and use them to form our objective functions. For convenience, the notation $\mathrm{diag}(M)$ henceforth refers to the diagonal matrix of the matrix M.

2.1. Fidelity Test. Considering A as the indexing matrix in the LSI, one way to infer whether Z is a good (but low-rank) representation of A is to use documents (rows) in A as queries themselves to inquire about document information from Z. If the document-term relationships embedded in Z adhere to those in A, then the scores of the query $\mathbf{q}_j$ applied to Z, where $\mathbf{q}_j^\top$ is the j-th row of A, should point to the most relevant document j. It should be so for each $j = 1, \ldots, n$. In other words, in the product $ZA^\top$ the highest score per column should occur precisely at the diagonal entry. This notion of self-examination is referred to as the *fidelity test* of Z to A.

Ideally, the fidelity test should product the maximum value 1 along the diagonal. Since the indexing matrix A is never perfect, we may as well be settled with the collective effect for the sake of easy computation. We say that the approximation Z remains in "good" fidelity to the original A if the trace of $ZA^\top$ is sufficiently close to $\mathrm{trace}(AA^\top) = n$. This low-rank approximation problem therefore becomes the problem of maximizing

$$\mathrm{trace}(ZA^\top) = \langle Z, A \rangle, \tag{2.1}$$

where $\langle M, N \rangle$ stands for the Frobenius inner product of matrices M and N in $\mathbb{R}^{n \times m}$. The objective, in short, is to *maximize* the fidelity of Z to A while maintaining its low rank.

In order to characterize the properties that the matrix Z is of rank k with unit-length rows, we introduce the parametrization that

$$Z = \left(\text{diag}\left(USS^\top U^\top\right)\right)^{-1/2} USV^\top, \tag{2.2}$$

where $U \in \mathcal{O}_n$, $S \in \mathbb{R}^k$, $V \in \mathcal{O}_m$, and $\mathcal{O}_n$ denotes the group of orthogonal matrices of size $n \times n$. The task now is equivalent to maximizing the functional

$$F(U,S,V) := \langle \left(\text{diag}\left(USS^\top U^\top\right)\right)^{-1/2} USV^\top, A\rangle, \tag{2.3}$$

subject to the orthogonal conditions of U and V. In the above, we have used the same symbol $S \in \mathbb{R}^k$ to denote, without causing ambiguity, the "diagonal matrix" in $\mathbb{R}^{n\times m}$ whose first k diagonal entries are those of S. Observe that by limiting S to $\mathbb{R}^k$, the rank of the matrix $X = USV^\top \in \mathbb{R}^{n\times m}$ is guaranteed to be at most k, and is exactly k if S contains no zero entry.

The Fréchet derivative of F at (U,S,V) acting on any $(H,D,K) \in \mathbb{R}^{n\times n} \times \mathbb{R}^k \times \mathbb{R}^{m\times m}$ can be considered as the sum of three operations,

$$F'(U,S,V).(H,D,K) = \frac{\partial F}{\partial U}.H + \frac{\partial F}{\partial S}.D + \frac{\partial F}{\partial V}.K, \tag{2.4}$$

where the operation $\Lambda.\eta$ denotes the result of the action by the linear operator Λ on the η. For convenience, define

$$\alpha(U,S) := \left(\text{diag}\left(USS^\top U^\top\right)\right)^{-1}. \tag{2.5}$$

We now calculate each action in (2.4) as follows.

First, it is easy to see that

$$\frac{\partial F}{\partial V}.K = \langle \alpha(U,S)^{1/2} USK^\top, A\rangle. \tag{2.6}$$

Using the property that

$$\langle MN, P\rangle = \langle M, PN^\top\rangle = \langle N, M^\top P\rangle \tag{2.7}$$

for any matrices M, N and P of compatible sizes, it follows from the Riesz representation theorem that, with respect to the Frobenius inner product, the partial gradient can be represented as

$$\frac{\partial F}{\partial V} = A^\top \alpha(U,S)^{1/2} US. \tag{2.8}$$

Observe next that

$$\frac{\partial F}{\partial S}.D = \langle \alpha(U,S)^{1/2} UDV^\top, A\rangle + \langle \left(\frac{\partial \alpha(U,S)^{1/2}}{\partial S}.D\right) USV^\top, A\rangle. \tag{2.9}$$

By the chain rule we have

$$\begin{aligned}\frac{\partial \alpha(U,S)^{1/2}}{\partial S}.D &= \frac{1}{2}\alpha(U,S)^{-1/2}\left(-\alpha(U,S)\text{diag}\left(UDS^\top U^\top + USD^\top U^\top\right)\alpha(U,S)\right)\\ &= -\frac{1}{2}\alpha(U,S)^{3/2}\text{diag}\left(UDS^\top U^\top + USD^\top U^\top\right),\end{aligned} \tag{2.10}$$

where we have used the fact that diagonal matrices are commutable. Denoting the diagonal matrix

$$\omega(U,S,V) := \alpha(U,S)^{3/2}\text{diag}\left(AVS^\top U^\top\right), \tag{2.11}$$

and taking advantage of the fact that $\langle M, \mathrm{diag}(N)\rangle = \langle \mathrm{diag}(M), \mathrm{diag}(N)\rangle$, the action in (2.9) can be expressed as

$$\begin{aligned}\frac{\partial F}{\partial S}.D &= \langle D, U^\top \alpha(U,S)^{1/2} AV\rangle - \frac{1}{2}\langle UDS^\top U^\top + USD^\top U^\top, \omega(U,S,V)\rangle \\ &= \langle D, \mathrm{diag}\left(U^\top \alpha(U,S)^{1/2} AV\right)\rangle - \langle D, \mathrm{diag}\left(U^\top \omega(U,S,V) US\right)\rangle. \end{aligned} \tag{2.12}$$

We thus conclude that

$$\frac{\partial F}{\partial S} = \mathrm{diag}_k\left(U^\top \alpha(U,S)^{1/2} AV\right) - \mathrm{diag}_k\left(U^\top \omega(U,S,V) US\right), \tag{2.13}$$

where, for clarity, we use diag_k to denote the first k diagonal elements and to emphasize that the partial gradient $\frac{\partial F}{\partial S}$ is in fact a vector in $\mathbb{R}^k$.

Finally,

$$\frac{\partial F}{\partial U}.H = \langle \alpha(U,S)^{1/2} HSV^\top, A\rangle + \langle \left(\frac{\partial \alpha(U,S)^{1/2}}{\partial U}.H\right) USV^\top, A\rangle, \tag{2.14}$$

whereas by chain rule again we know that

$$\begin{aligned}\frac{\partial \alpha(U,S)^{1/2}}{\partial U}.H &= \frac{1}{2}\alpha(U,S)^{-1/2}\left(-\alpha(U,S)\mathrm{diag}\left(HSS^\top U^\top + USD^\top H^\top\right)\alpha(U,S)\right) \\ &= -\frac{1}{2}\alpha(U,S)^{3/2}\mathrm{diag}\left(HSS^\top U^\top + USS^\top H^\top\right). \end{aligned} \tag{2.15}$$

Together, we obtain

$$\begin{aligned}\frac{\partial F}{\partial U}.H &= \langle H, \alpha(U,S)^{1/2} AVS^\top\rangle - \frac{1}{2}\langle HSS^\top U^\top + USS^\top H^\top, \omega(U,S,V)\rangle \\ &= \langle H, \alpha(U,S)^{1/2} AVS^\top\rangle - \langle H, \omega(U,S,V) USS^\top\rangle, \end{aligned} \tag{2.16}$$

and hence

$$\frac{\partial F}{\partial U} = \alpha(U,S)^{1/2} AVS^\top - \omega(U,S,V) USS^\top. \tag{2.17}$$

Formulas (2.8), (2.13), and (2.17) constitute the gradient

$$\nabla F(U,S,V) = \left\langle \frac{\partial F}{\partial U}, \frac{\partial F}{\partial S}, \frac{\partial F}{\partial V}\right\rangle$$

of the objection function F. Because our parameters (U,S,V) must come from the manifold $\mathcal{O}_n \times \mathbb{R}^k \times \mathcal{O}_m$, we want to know further the projection of ∇F into the feasible set.

By taking advantage of the product topology, the tangent space $\mathcal{T}_{(U,S,V)}(\mathcal{O}_n \times \mathbb{R}^k \times \mathcal{O}_m)$ of the product manifold $\mathcal{O}_n \times \mathbb{R}^k \times \mathcal{O}_m$ at $(U,S,V) \in \mathcal{O}_n \times \mathbb{R}^k \times \mathcal{O}_m$ can be decomposed as the product of tangent spaces, i.e.,

$$\mathcal{T}_{(U,S,V)}(\mathcal{O}_n \times \mathbb{R}^k \times \mathcal{O}_m) = \mathcal{T}_U\mathcal{O}_n \times \mathbb{R}^k \times \mathcal{T}_V\mathcal{O}_m. \tag{2.18}$$

The projection of $\nabla F(U,S,V)$ onto $\mathcal{T}_{(U,S,V)}(\mathcal{O}_n \times \mathbb{R}^k \times \mathcal{O}_m)$, therefore, is the product of the projections of $\frac{\partial F}{\partial U}$, $\frac{\partial F}{\partial S}$, and $\frac{\partial F}{\partial V}$ onto $\mathcal{T}_U\mathcal{O}_n$, $\mathbb{R}^k$, and $\mathcal{T}_V\mathcal{O}_m$, respectively. Each of the projections can easily be calculated by using the techniques developed in [8] which we summarize below.

Observe that any matrix $M \in \mathbb{R}^{n\times n}$ has a unique orthogonal splitting,

$$M = U\left\{\frac{1}{2}(U^\top M - M^\top U)\right\} + U\left\{\frac{1}{2}(U^\top M + M^\top U)\right\}, \tag{2.19}$$

as the sum of elements from the tangent space $\mathcal{T}_U O_n$ and the normal space $\mathcal{N}_U\mathcal{O}_n$. Thus the projection $\mathcal{P}_{\mathcal{T}_U\mathcal{O}_n}(M)$ of M onto $\mathcal{T}_U\mathcal{O}_n$ is given by the matrix

$$\mathcal{P}_{\mathcal{T}_U\mathcal{O}_n}(M) = U\left\{\frac{1}{2}(U^\top M - M^\top U)\right\}. \tag{2.20}$$

Replacing M by $\frac{\partial F}{\partial U}$ in (2.20), we thus obtain the explicit formulation of the projection of $\frac{\partial F}{\partial U}$ onto the tangent space $\mathcal{T}_U O_n$, that is,

$$\mathcal{P}_{\mathcal{T}_U\mathcal{O}_n}(\frac{\partial F}{\partial U}) = \Big\{\big(XX^T\omega(U,S,V) - \omega(U,S,V)XX^T\big) + \Big(\alpha(U,S)^{1/2}AX^T - XA^T\alpha(U,S)^{1/2}\Big)\Big\}\,\frac{U}{2}.$$

Similarly, the projection of $\frac{\partial F}{\partial V}$ onto tangent spaces $\mathcal{T}_V\mathcal{O}_m$ can be calculated. The projection of $\frac{\partial F}{\partial S}$ onto $\mathbb{R}^k$ is just itself. These projections characterize precisely the projected gradient of the objective function $F(U,S,V)$ on the manifold $\mathcal{O}_n \times \mathbb{R}^k \times \mathcal{O}_m$.

The explicit formulation of the project gradient serves as the first order optimality condition that any local solution must satisfy. Many iterative methods making use of projected gradient for constrained optimization are available in the literature [11, 19]. On the other hand, we find it most natural and convenient to use the dynamical system,

$$\begin{aligned}
\frac{dU}{dt} &= \mathcal{P}_{\mathcal{T}_U\mathcal{O}_n}(\frac{\partial F}{\partial U}),\\
\frac{dS}{dt} &= \frac{\partial F}{\partial S},\\
\frac{dV}{dt} &= \mathcal{P}_{\mathcal{T}_V\mathcal{O}_m}(\frac{\partial F}{\partial V}),
\end{aligned} \tag{2.21}$$

to navigate a flow on the manifold $\mathcal{O}_n \times \mathbb{R}^k \times \mathcal{O}_m$. The flow, called the *fidelity flow* for later reference, moves along the steepest ascent direction to maximize the objective functional F. Since (2.21) defines an ascent flow by an analytic vector field, by the well-known Łojasiewicz-Simon theorem [6, 25, 30] the flow converges to a single point (U_*, S_*, V_*) of equilibrium at which

$$Z_* = \left(\operatorname{diag}\left(U_*S_*S_*^\top U_*^\top\right)\right)^{-1/2} U_*S_*V_*^\top \tag{2.22}$$

is locally the best rank k approximation to A subject to constraint of unit row vectors.

2.2. Nearness Test. In the context of LSI, each document is represented by a single point in the term space $\mathbb{R}^m$. It is natural to evaluate the quality of an approximation Z to A by measuring how far Z is from A. The low-rank approximation problem therefore becomes the problem of *minimizing* the functional

$$E(U,S,V) := \frac{1}{2}\langle \alpha(U,S)^{1/2}USV^\top - A, \alpha(U,S)^{1/2}USV^\top - A\rangle, \tag{2.23}$$

subject to the constraints that $U \in \mathcal{O}_n$, $S \in \mathbb{R}^k$, and $V \in \mathcal{O}_m$. This notion of finding low-rank matrices nearest to the given A is called the *nearest test*. Nearest low-rank approximation problems under different (linear) structures have been considered in [9, 14, 27]. By a linear structure, we mean that all matrices of the same structure form an affine subspace. Our structure, defined by the requirement that each row is of unit length, is not a linear structure.

We quickly point out the fact that

$$\frac{1}{2}\langle Z-A, Z-A\rangle = \frac{1}{2}(\langle Z,Z\rangle + \langle A,A\rangle) - \langle Z,A\rangle. \tag{2.24}$$

Note that $\langle Z,Z\rangle = \langle A,A\rangle = n$. Therefore, minimizing the functional $E(U,S,V)$ in (2.23) is equivalent to maximizing the functional $F(U,S,V)$ in (2.3). The fidelity test is exactly the same as the nearest test under the Frobenius inner product.

2.3. Absolute Fidelity Test. The data matrix A in general might contain negative entries. In this case, cancellations among terms in the trace calculation might inadvertently reduce the fidelity quality. To prevent cancellation, we might want to consider *maximizing* the sum of squares of the diagonal elements of $ZA^\top$. The objective functional to be maximized is

$$G(U,S,V) := \frac{1}{2}\langle \mathrm{diag}\left(\alpha(U,S)^{1/2}USV^\top A^\top\right), \mathrm{diag}\left(\alpha(U,S)^{1/2}USV^\top A^\top\right)\rangle, \tag{2.25}$$

subject to the constraints that $U \in \mathcal{O}_n$, $S \in \mathbb{R}^k$, and $V \in \mathcal{O}_m$. We remark in passing that in a certain sense the absolute fidelity test has the effect of rewarding both anti-correlation and positive correlation between matrices A and Z. In certain applications, such as identifying genes with similar functionalities, detecting anti-correlated gene clusters sometimes is as important as detecting positively correlated genes [15].

Once again, denote the Fréchet derivative of G at (U,S,V) acting on any $(H,D,K) \in \mathbb{R}^{n\times n} \times \mathbb{R}^k \times \mathbb{R}^{m\times m}$ as

$$G'(U,S,V).(H,D,K) = \frac{\partial G}{\partial U}.H + \frac{\partial G}{\partial S}.D + \frac{\partial G}{\partial V}.K. \tag{2.26}$$

We can calculate each partial derivative in (2.26) to obtain the gradient ∇G.

For convenience, denote

$$\theta(U,S,V) := \mathrm{diag}\left(\alpha(U,S)^{1/2}USV^\top A^\top\right). \tag{2.27}$$

Observe that $\mathrm{diag}(\cdot)$ is a linear operator, so its derivative is just itself. It follows that

$$\frac{\partial G}{\partial V}.K = \langle \alpha(U,S)^{1/2}USK^\top A^\top, \theta(U,S,V)\rangle. \tag{2.28}$$

Using (2.7) and the fact that diagonal matrices commute, we see that

$$\frac{\partial G}{\partial V} = A^\top \theta(U,S,V)^\top \alpha(U,S)^{1/2}US = A^\top \alpha(U,S)\mathrm{diag}\left(USV^\top A^\top\right)US. \tag{2.29}$$

Likewise,

$$\begin{aligned}\frac{\partial G}{\partial S}.D = &\langle \mathrm{diag}\left(\alpha(U,S)^{1/2}UDV^\top A^\top\right), \theta(U,S,V)\rangle + \\ &+\langle \mathrm{diag}\left(\left(\frac{\partial \alpha(U,S)^{1/2}}{\partial S}.D\right)USV^\top A^\top\right), \theta(U,S,V)\rangle,\end{aligned}$$

whereas the action $\frac{\partial \alpha(U,S)^{1/2}}{\partial S}.D$ is already calculated in (2.10). Upon substitution, we obtain

$$\begin{aligned}\frac{\partial G}{\partial S}.D &= \langle D, \mathrm{diag}\left(U^\top \alpha(U,S)^{1/2}\theta(U,S,V)AV\right)\rangle \\ &\quad -\frac{1}{2}\langle \mathrm{diag}\left(UDS^\top U^\top + USD^\top U^\top\right), \alpha(U,S)^{3/2}\theta(U,S,V)\mathrm{diag}\left(AVS^\top U^\top\right)\rangle \\ &= \langle D, \mathrm{diag}\left(U^\top \alpha(U,S)^{1/2}\theta(U,S,V)AV\right)\rangle \\ &\quad -\langle D, \mathrm{diag}\left(U^\top \alpha(U,S)^{3/2}\theta(U,S,V)\mathrm{diag}(AVS^\top U^\top)US\right)\rangle.\end{aligned} \tag{2.30}$$

We thus conclude that

$$\frac{\partial G}{\partial S} = \mathrm{diag}_k\left(U^\top \alpha(U,S)^{1/2}\theta(U,S,V)AV\right) \\ -\mathrm{diag}_k\left(U^\top \alpha(U,S)^{3/2}\theta(U,S,V)\mathrm{diag}\left(AVS^\top U^\top\right)US\right). \tag{2.31}$$

Finally,

$$\frac{\partial G}{\partial U}.H = \langle \mathrm{diag}\left(\alpha(U,S)^{1/2}HSV^\top A^\top\right), \theta(U,S,V)\rangle + \\ \langle \mathrm{diag}\left(\left(\frac{\partial \alpha(U,S)^{1/2}}{\partial U}.H\right)USV^\top A^\top\right), \theta(U,S,V)\rangle.$$

Substituting the expression of $\frac{\partial \alpha(U,S)^{1/2}}{\partial U}.H$ computed in (2.15) and simplifying, we see that

$$\begin{aligned}\frac{\partial G}{\partial U}.H &= \langle H, \alpha(U,S)^{1/2}\theta(U,S,V)AVS^\top\rangle \\ &\quad -\frac{1}{2}\langle \mathrm{diag}\left(HSS^\top U^\top + USS^\top H^\top\right), \alpha(U,S)^{3/2}\theta(U,S,V)AVS^\top U^\top\rangle \\ &= \langle H, \alpha(U,S)^{1/2}\theta(U,S,V)AVS^\top\rangle \\ &\quad -\langle H, \alpha(U,S)^{3/2}\theta(U,S,V)\mathrm{diag}\left(AVS^\top U^\top\right)USS^\top\rangle\end{aligned}$$

and obtain

$$\frac{\partial G}{\partial U} = \alpha(U,S)^{1/2}\theta(U,S,V)AVS^\top - \alpha(U,S)^{3/2}\theta(U,S,V)\mathrm{diag}\left(AVS^\top U^\top\right)USS^\top). \tag{2.32}$$

Formulas (2.29), (2.31), and (2.32) constitute the gradient,

$$\nabla G = \left\langle \frac{\partial G}{\partial U}, \frac{\partial G}{\partial S}, \frac{\partial G}{\partial V}\right\rangle,$$

of the objection function G in the general ambient space $\mathbb{R}^{n\times n}\times\mathbb{R}^k\times\mathbb{R}^{m\times m}$. Similar to the preceding section, we can express by using (2.20) the projected gradient of G in explicit form to establish the first order optimality condition. We can also formulate the *absolute fidelity flow* that evolves on the manifold $\mathcal{O}_n\times\mathbb{R}^k\times\mathcal{O}_m$ to maximize $G(U,S,V)$.

We conclude this section by pointing out that the formulas of the vector fields discussed above are not as computationally extensive as they appear to be. Many blocks of expressions repeat themselves in the formulas whereas all radicals of matrices acquired involve only diagonal matrices.

3. Compact Form and Stiefel Manifold. The above discussion provides a basis for computing the low-rank approximation. We must notice, however, that the square matrices are employed just for the sake of deriving the formulas. A lot of information computed is not needed in defining Z in (2.2). It suffices to know only a portion of columns of $U\in\mathcal{O}_n$ and $V\in\mathcal{O}_m$, particularly when the desirable rank k of Z is assumed to be much smaller than m and n. Indeed, we may write Z precisely in the same way as (2.2), but assume that

$$U\in\mathcal{O}(n,k),\quad S\in\mathbb{R}^k,\quad \text{and } V\in\mathcal{O}(m,k), \tag{3.1}$$

where

$$\mathcal{O}(p,q) := \{Q\in\mathbb{R}^{p\times q} | Q^\top Q = I_q\} \tag{3.2}$$

denotes the set of all $p \times q$ real matrices with orthonormal columns. This set, known as the Stiefel manifold, forms a smooth manifold [32].

The set $\mathcal{O}(p,q)$ enjoys many properties similar to those of the orthogonal group. Upon close examination, it can be checked that all expressions derived above for the gradients of $F(U,S,V)$ and $G(U,S,V)$ remain valid, even if we restrict U to $\mathcal{O}(n,k)$ and V to $\mathcal{O}(m,k)$. It only remains to compute the projected gradients onto the Stiefel manifolds. We outline in this section some main points for this purpose.

Embedding $\mathcal{O}(p,q)$ in the Euclidean space $\mathbb{R}^{p\times q}$ which is equipped with the Frobenius inner product, it is easy to see that any vector H in the tangent space $\mathcal{T}_Q\mathcal{O}(p,q)$ is necessarily of the form

$$H = QK + (I_p - QQ^\top)W, \tag{3.3}$$

where $K \in \mathbb{R}^{q\times q}$ and $W \in \mathbb{R}^{p\times q}$ are arbitrary, and K is skew-symmetric. Furthermore, the space $\mathbb{R}^{p\times q}$ can be written as the direct sum of three mutually perpendicular subspaces

$$\mathbb{R}^{p\times q} = Q\mathcal{S}(q) \oplus \mathcal{N}(Q^T) \oplus Q\mathcal{S}(q)^\perp, \tag{3.4}$$

where $\mathcal{S}(q)$ is subspace of $q\times q$ symmetric matrices, $\mathcal{S}(q)^\perp$ is the subspace of $q\times q$ skew-symmetric matrices, and $\mathcal{N}(Q^\top) := \{X \in \mathbb{R}^{p\times q} | Q^\top X = 0\}$. Any $M \in \mathbb{R}^{p\times q}$ can be uniquely split as

$$M = Q\frac{Q^\top M - M^\top Q}{2} + (I - QQ^\top)M + Q\frac{Q^\top M + M^\top Q}{2}. \tag{3.5}$$

Similar to (2.20), it follows that the projection $\mathcal{P}_{\mathcal{O}(p,q)}(M)$ of $M \in \mathbb{R}^{p\times q}$ onto the tangent space $\mathcal{T}_Q\mathcal{O}(p,q)$ is given by

$$\mathcal{P}_{\mathcal{O}(p,q)}(M) = Q\frac{Q^\top M - M^\top Q}{2} + (I - QQ^\top)M. \tag{3.6}$$

Note that in case $p = q$ so that Q is orthogonal, the second term in (3.6) is identically zero and the above notion is identical to that of (2.20).

We may now project each of the partial gradients calculated in the preceding sections onto the appropriate Stiefel subspace according to (3.6) and obtain projected gradient flows. The gain is that the dimensionality is much reduced and the computation is much more cost efficient. For a given $A \in \mathbb{R}^{n\times m}$ and a desirable rank k, the gradient flow on the orthogonal group would involve $n^2 + k + m^2$ dependent variables whereas the flow on the Stiefel manifold involves only $(n + m + 1)k$ variables.

Be cautious that because $\mathcal{O}(m,k)$ is a subspace of $\mathcal{O}_m$, the projections onto $\mathcal{O}(m,k)$ and $\mathcal{O}_m$ are different in general. The resulting dynamics of the gradient flows on the manifolds will behave differently. For computation, we prefer the compacted flows on the Stiefel manifolds.

4. Numerical Experiments. The information of projected gradients can be utilized in various ways to tackle the low-rank approximation problem. See, for example, the review article [20] where a variety of Lie group methods, continuous or discrete, have been discussed. For demonstration purpose, we report in this section some preliminary numerical results from using the gradient flow approach only. To make a distinction between the two gradient flows related to (2.3) and (2.25), we further denote the corresponding solutions by Z_F and Z_G, respectively. For convenience, we shall employ existing routines in Matlab as the ODE integrators, even though more elaborate coding, especially the recently developed geometric integrators techniques, might be more effective. At the moment, our primary concern is to focus more on the dynamical behavior of the resulting flows than on the efficiency. Lots more can be done to improve the efficiency.

The ODE Suite [29] in Matlab contains several variable step-size solvers at our disposal. Considering the fact that the original data matrix A is not precise in its own right in practice, high accuracy approximation of A is not needed. We set both local tolerance $AbsTol = RelTol = 10^{-6}$ while maintaining all other parameters at the default values of the Matlab codes.

Example 1. To visualize the effect of data fitting by great circles, consider the case where $A \in \mathbb{R}^{n\times 3}$. These points on S^2 can be conveniently represented in $\mathbb{R}^2$ by their azimuths θ and elevations ϕ. A rank 2 approximation of A corresponds to points on the intersection of S^2 with a plane passing through the origin. The fidelity test is equivalent to finding the nearest great circle to the original points on S^2 in the sense of least squares.

In the left drawing of Figure 4.1, we randomly generate $n = 120$ uniformly distributed angles $0 \le \theta \le 2\pi$ and $0 \le \phi \le \pi$. These angles are used as azimuths and elevations to produce the 120×3 matrix A. A total of 248 independent variables are involved in the dynamical system (2.21) to calculate the rank 2 approximation matrix Z_F which then is converted into azimuths and elevations. The resulting more regular pattern (sinuous curve) of the angles for Z_F in the left drawing of Figure 4.1 demonstrates the great circle approximation.

Suppose we limit the elevations to be in between the cones $\frac{\pi}{6} \le \phi \le \frac{\pi}{4}$. These rows of A reside within a circular band in the northern hemisphere. It would be difficult to predict how a great circle would fit these data. In the right drawing of Figure 4.1, it is interesting to observe that the azimuths θ on the best great circle approximation are distributed over only half of a circle.

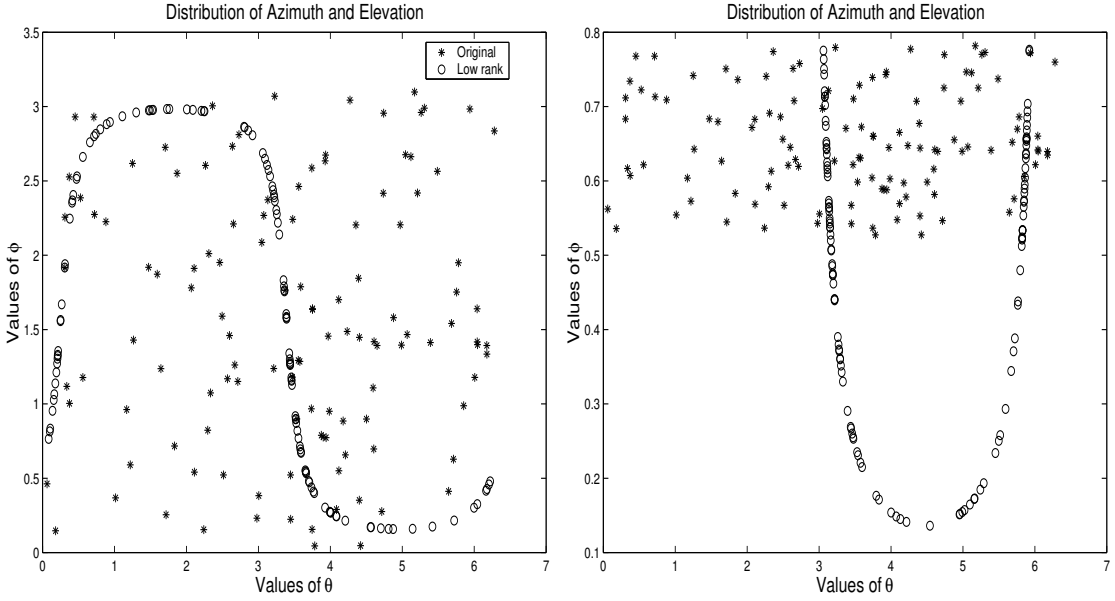

FIG. 4.1. *Big circle approximation of points on the unit sphere* S^2.

Example 2. Given a matrix A and a desirable low rank k, it is known that the (globally) nearest approximation of rank k to A is given by the truncated SVD $A^{(k)}$ of A. However, $A^{(k)}$ generally does not satisfy the normalization condition. Let $Z_0^{(k)}$ denote the row-wise normalized $A^{(k)}$. This example demonstrates that the normalized TSVD $Z_0^{(k)}$ generally does not maintain the highest fidelity.

Let $Z_*^{(k)}$ denote the (nearly) limit point of the fidelity flow using $Z_0^{(k)}$ as the starting value.

rank k	$\|A^{(k)} - A\|_2$	$\|A^{(k)} - A\|_F$	$\|Z_0^{(k)} - A\|_F$	$\|Z_*^{(k)} - A\|_F$	$\|Z_0^{(k)} - Z_*^{(k)}\|_F$
1	1.5521	2.7426	3.4472	3.3945	0.9227
2	1.4603	2.2612	2.6551	2.5330	2.1866
3	1.3056	1.7264	1.9086	1.8588	0.8085
4	1.1291	1.1296	1.1888	1.1744	0.3341
5	0.0237	0.0343	0.0343	0.0343	5.5969×10^{-7}
6	0.0183	0.0248	0.0248	0.0248	2.1831×10^{-7}
7	0.0141	0.0167	0.0167	0.0167	1.1083×10^{-7}
8	0.0083	0.0089	0.0089	0.0089	6.4180×10^{-9}
9	0.0034	0.0034	0.0034	0.0034	$2,7252 \times 10^{-9}$

TABLE 4.1
Comparison of nearness of TSVD and fidelity flow

Table 4.1 compares the distances $\|A^{(k)} - A\|_F$, $\|Z_0^{(k)} - A\|_F$, $\|Z_*^{(k)} - A\|_F$, and $\|Z_0^{(k)} - Z_*^{(k)}\|_F$ of a 10×10 random matrix A with controlled singular values. For reference, we have also listed $\|A^{(k)} - A\|_2$ which is known to be the $(k+1)$-th singular value of A.

Because $\sum_{i=1}^{10} \sigma_i^2 = 10$, we purposefully control the singular values of A so that five singular values are relatively larger than the other five. We then notice that the gap $\|Z_0^{(k)} - Z_*^{(k)}\|_F$ is relatively significant up to $k = 4$. Since the local tolerance in the ODE integrator is set only at 10^{-6}, the difference between $Z_0^{(k)}$ and $Z_*^{(k)}$ is probably negligible when $k \geq 5$.

Example 3. We measure the fidelity and the absolute fidelity by the ratios

$$r(Z_F) = \frac{\text{trace}(Z_F A^\top)}{\text{trace}(AA^\top)} = \frac{\text{trace}(Z_F A^\top)}{n}, \tag{4.1}$$

$$q(Z_G) = \frac{\|\text{diag}(Z_G A^\top)\|_2^2}{\|\text{diag}(AA^\top)\|_2^2} = \frac{\|\text{diag}(Z_G A^\top)\|_2^2}{n}, \tag{4.2}$$

respectively. The highest attainable value for either $r(Z_F)$ or $q(Z_G)$ should be one.

The left drawing in Figure 4.2 represents a typical attainable $r(Z_F)$ of a rank k approximation Z_F to a matrix A of size 120×50. For each $k = 2, \ldots, 50$, we are solving a dynamical system of $171k$ many independent variables. These are fairly large-scale dynamical systems when k is large. Our proposed projected gradient flow approach thus connects to and can benefit from another interesting research subject — model reduction. See, for example, [1, 5] for an overview and many other studies in the literature. The right drawing in Figure 4.2 compares the attainable $r(Z_F)$ and $q(Z_G)$ to a matrix A of size 200×50 with rank $k = 2, \ldots, 14$. The plots in Figure 4.2 seem to suggest that the value of fidelity is an increasing function of k. However, such an observation is true only because we systematically modify the initial values one column a time. The different scales along the y-axis in the left and right drawings clearly illustrate the effect of different initial values for the flows. The initial values for experiments involved in the left drawing are randomly generated and are very far away from A, while those for experiments in the right drawing are built to have at least 80% fidelity to A to begin with.

5. Conclusion and Future Work. Motivated by the fact that in many applications data in high dimensional space are normalized to unit length, this paper touches upon the issue of best approximation by low dimensional data while maintaining unit length. The problem is cast as a low-rank matrix approximation subject to row normalization. After proper parametrization of the feasible set, we offer explicit formulas for the project gradients which can be turned into a numerical means for computing the approximation.

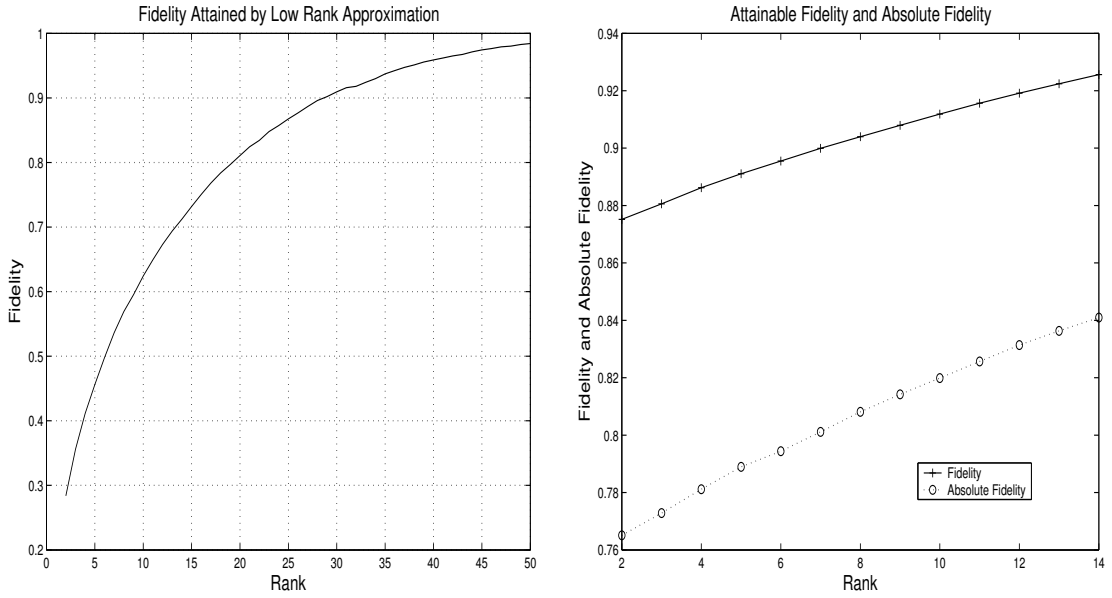

FIG. 4.2. *Attainable fidelity with various rank values and initial values. Left:* $A \in \mathbb{R}^{120\times 50}$ *and random initial values. Right:* $A \in \mathbb{R}^{200\times 50}$ *and high fidelity initial values.*

We experiment with a projected gradient flow approach. The capacity of the differential equation approach is limited only by current development of numerical ODE techniques. Thus far, a direct ODE approach can handle matrices of sizes up to thousands within a reasonable frame of time, say, a few CPU minutes in a desktop computing environment. However, it may not be efficient enough to handle very large-scale problem which often is the case in real applications. Other techniques such as model reduction might prove beneficial, but are not discussed in the presentation. The availability of projected gradient in explicit form also makes it possible to apply other optimization techniques including iterative methods. It might be worth further investigation in this direction. It is our hope that this study will have shed some light on the interesting low-rank approximation problem.

REFERENCES

[1] A. C. Antoulas and D. C. Sorensen, Approximation of large-scale dynamical systems: an overview, Int. J. Appl. Math. Comput. Sci., 11(2001), 1093-1121.

[2] M. W. Berry, Computational Information Retrieval, SIAM, Philadelphia, 2000.

[3] R. De Beer, Quantitative in vivo NMR (Nuclear Magnetic Resonance on Living Objects), available at *http://dutnsic.tn.tudelft.nl:8080/c59_to_html/c59.html*, University of Technology Delft, Netherlands, 1995.

[4] J. P. Burg, D. G. Luenberger, and D. L. Wenger [1982], Estimation of structured covariance Matrices, Proceedings of the IEEE, 70(1982), 963–974.

[5] Y. Chahlaoui, Low-rank approximation and model reduction, Ph.D. Thesis, Département d'ingénierie mathématique, Université Catholique De Louvain, 2003.

[6] R. Chill, On the Łojasiewicz-Simon gradient inequality, J. Func. Anal., 201(2003), 572-601.

[7] M. T. Chu, On the statistical meaning of the truncated singular decomposition, preprint, North Carolina State University, November, 2000.

[8] M. T. Chu and K. R. Driessel, The projected gradient method for least squares matrix approximations with spectral constraints, SIAM J. Numer. Anal. 27(1990), 1050-1060.

[9] M. T. Chu, R. E. Funderlic, and R. J. Plemmons, Structured low-rank approximation, Linear Alg. Appl., 366(2003), 157-172.

[10] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, Indexing by latent semantic analysis, J. Amer. Soc. Inform. Sci., 41(1990), 391-407.
[11] P. E. Gill, W. Murray, and M. H. Wright, Practical Optimization, Academic Press, London, 1981.
[12] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, Wiley, New York, 2001.
[13] M. Dendrinos, S. Bakamidis, and G. Carayannis, Speech enhancement from noise: A regenerative approach, Speech Communication 10(45-57), 1991.
[14] I. S. Dhillon and D. M. Modha, Concept decompositions for large sparse text data using clustering, Machine Learning J., 42(2001), 143-175.
[15] I. S. Dhillon, E. M. Marcotte, and U. Roshan, Diametrical clustering for identifying anti-correlated gene clusters, Bioinformatics, 19(2003), 1612-1619.
[16] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, New York, 2001.
[17] P. Horst, Factor Analysis of Data Matrices, Holt, Rinehart and Winston, New York, 1965.
[18] L. Hubert, J. Meulman and Willem Heiser, Two purposes for matrix factorization: A historical appraisal, SIAM Review, 42(2000), 68-82.
[19] A. N. Iusem, On the convergence properties of the projected gradient method for convex optimization, Comput. Appl. Math., 22(2003), 37-52.
[20] A. Iserles, H. Z. Munthe-Kass, S. P. Nørsett, and A. Zanna, Lie-group methods, Acta Numerica, 2000, 215-365.
[21] I. T. Jolliffe, Principal Component Analysis, 2nd ed., Springer-Verlag, New York, 2002.
[22] J. Kleinberg, C. Papadimitriou, and P. Raghavan, A microeconomic view of data mining, Data Mining and Knowledge Discovery, 2(1998), 311-324.
[23] T. G. Kolda and D. P. O'Leary, A semi-discrete matrix decomposition for latent semantic indexing in information retrieval, ACM Transact. Inform. Systems, 16(1998), 322-346.
[24] B. De Moor, Total least squares for affinely structured matrices and the noisy realization problem, IEEE Trans. Signal Processing, 42(1994), 3104-3113.
[25] S. Łojasiewicz, Une propriété topologique des sous-ensembles analytiques réels, in Les Équations aux Dérivées Partielles (Paris, 1962), Vol. 117, 87-89, Éditions du Centre National de la Recherche Scientifique, Paris, 1963.
[26] P. Paatero, Least squares formulation of robust nonnegative factor analysis, Chemomet. Intell. Lab. Systems, 37(1997), 23-35.
[27] H. Park, M. Jeon, and J. B. Rosen, Lower dimensional representation of text data in vector space based information retrieval, in Computational Information Retrieval, ed. M. Berry, Proc. Comput. Inform. Retrieval Conf., SIAM, 2001, 3-23.
[28] G. Salton and M. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.
[29] L. F. Shampine and M. W. Reichelt, The MATLAB ODE suite, SIAM J. Sci. Comput., 18(1997), 1-22.
[30] L. Simon, Asymptotics for a class of nonlinear evolution equations with applications to geometric problems, Annals Math., 118(1983), 525-571.
[31] A. Singhal, C. Buckley, and M. Mitra, Pivoted document lenght normalization, in Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrival, 1996, 21-29.
[32] E. Stiefel, Richtungsfelder und fernparallelismus in n-dimensionalel manning faltigkeiten, Commentarii Mathematici Helvetici, 8(1935-1936), 305-353.
[33] P. Stoica and M. Viberg, Maximum likelihood parameter and rank esitmation in reduced rank multivariate linear regressions, IEEE Trans. Signal Process, 44(1996), 3069-3078.
[34] J. Tropp,, I. Dhillon, R. Heath, T. Strohmer, Designing structured tight frames via an alternating projection method, IEEE Trans. Info. Theory, to appear, 2004.