Orthogonal Nonnegative Matrix Factorization: Multiplicative Updates on Stiefel Manifolds

Jiho Yoo and Seungjin Choi

Department of Computer Science
Pohang University of Science and Technology
San 31 Hyoja-dong, Nam-gu, Pohang 790-784, Korea
{zentasis,seungjin}@postech.ac.kr

Abstract. Nonnegative matrix factorization (NMF) is a popular method for multivariate analysis of nonnegative data, the goal of which is decompose a data matrix into a product of two factor matrices with all entries in factor matrices restricted to be nonnegative. NMF was shown to be useful in a task of clustering (especially document clustering). In this paper we present an algorithm for orthogonal nonnegative matrix factorization, where an orthogonality constraint is imposed on the nonnegative decomposition of a term-document matrix. We develop multiplicative updates directly from true gradient on Stiefel manifold, whereas existing algorithms consider additive orthogonality constraints. Experiments on several different document data sets show our orthogonal NMF algorithms perform better in a task of clustering, compared to the standard NMF and an existing orthogonal NMF.

1 Introduction

Nonnegative matrix factorization (NMF) is a multivariate analysis method which is proven to be useful in learning a faithful representation of nonnegative data such as images, spectrograms, and documents [1]. NMF seeks a decomposition of a nonnegative data matrix into a product of basis and encoding matrices with all of these matrices restricted to have only nonnegative elements. NMF allows only non-subtractive combinations of nonnegative basis vectors to approximate the original nonnegative data, possibly providing a parts-based representation [1]. Incorporating extra constraints such as locality and orthogonality was shown to improve the decomposition, identifying better local features or providing more sparse representation [2]. Orthogonality constraints were imposed on NMF [3], where nice clustering interpretation was studied in the framework of NMF.

A prominent application of NMF is in document clustering [4, 5], where a decomposition of a term-document matrix was considered. In this paper we consider *orthogonal NMF* and its application to document clustering, where an orthogonality constraint is imposed on the nonnegative decomposition of a term-document matrix. We develop new multiplicative updates for orthogonal NMF, which are directly derived from true gradient on Stiefel manifold, while existing algorithms consider additive orthogonality constraints. Experiments on several

different document data sets show our orthogonal NMF algorithms perform better in a task of clustering, compared to the standard NMF and an existing orthogonal NMF.

2 NMF for document clustering

In the vector-space model of text data, each document is represented by an m-dimensional vector $\boldsymbol{x}_t \in \mathbb{R}^m$, where m is the number of terms in the dictionary. Given N documents, we construct a term-document matrix $\boldsymbol{X} \in \mathbb{R}^{m \times N}$ where X_{ij} corresponds to the significance of term t_i in document d_j that is calculated by

$$X_{ij} = \mathrm{TF}_{ij} \log \left(\frac{N}{\mathrm{DF}_i} \right),$$

where TF_{ij} denotes the frequency of term t_i in document d_j and DF_i represents the number of documents containing term t_i. Elements X_{ij} are always nonnegative and equal zero only when corresponding terms do not appear in the document.

NMF seeks a decomposition of $\boldsymbol{X} \in \mathbb{R}^{m \times N}$ that is of the form

$$\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{V}^\top, \tag{1}$$

where $\boldsymbol{U} \in \mathbb{R}^{m \times K}$ and $\boldsymbol{V} \in \mathbb{R}^{N \times K}$ are restricted to be nonnegative matrices as well and K corresponds to the number of clusters when NMF is used for clustering. Matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, in general, are interpreted as follows.

- When columns in $\boldsymbol{X}$ are treated as data points in m-dimensional space, columns in $\boldsymbol{U}$ are considered as *basis vectors* (or *factor loadings*) and each row in $\boldsymbol{V}$ is *encoding* that represents the extent to which each basis vector is used to reconstruct each data vector.
- Alternatively, when rows in $\boldsymbol{X}$ are data points in N-dimensional space, columns in $\boldsymbol{V}$ correspond to basis vectors and each row in $\boldsymbol{U}$ represents encoding.

Applying NMF to a term-document matrix for document clustering, each column of $\boldsymbol{X}$ is treated as a data point in m-dimensional space. In such a case, the factorization (1) is interpreted as follows.

- U_{ij} corresponds to the degree to which term t_i belongs to cluster c_j. In other words column j of $\boldsymbol{U}$, denoted by $\boldsymbol{u}_j$, is associated with a prototype vector (center) for cluster c_j.
- V_{ij} corresponds to the degree document d_i is associated with cluster j. With appropriate normalization, V_{ij} is proportional to a posterior probability of cluster c_j given document d_i. More details on probabilistic interpretation of NMF for document clustering are summarized in Sec. 2.2.

2.1 Multiplicative updates for NMF

We consider the squared Euclidean distance as a discrepancy measure between the data $\boldsymbol{X}$ and the model $\boldsymbol{U}\boldsymbol{V}^\top$, leading to the following least squares error function

$$\mathcal{E} = \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^\top\|^2. \tag{2}$$

NMF involves the following optimization:

$$\underset{U\geq 0, V\geq 0}{\arg\min}\ \mathcal{E} = \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^\top\|^2. \tag{3}$$

Gradient descent learning (which is additive update) can be applied to determine a solution to (3), however, nonnegativity for $\boldsymbol{U}$ and $\boldsymbol{V}$ is not preserved without further operations at iterations.

On the other hand, a multiplicative method developed in [6] provides a simple algorithm for (3). We give a slightly different approach from [6] to derive the same multiplicative algorithm. Suppose that the gradient of an error function has a decomposition that is of the form

$$\nabla\mathcal{E} = [\nabla\mathcal{E}]^+ - [\nabla\mathcal{E}]^-, \tag{4}$$

where $[\nabla\mathcal{E}]^+ > 0$ and $[\nabla\mathcal{E}]^- > 0$. Then multiplicative update for parameters Θ has the form

$$\Theta \leftarrow \Theta \odot \left(\frac{[\nabla\mathcal{E}]^-}{[\nabla\mathcal{E}]^+}\right)^{\cdot\eta}, \tag{5}$$

where $\odot$ represents Hadamard product (elementwise product) and $(\cdot)^{\cdot\eta}$ denotes the elementwise power and η is a learning rate ($0 < \eta \leq 1$). It can be easily seen that the multiplicative update (5) preserves the nonnegativity of the parameter Θ, while $\nabla\mathcal{E} = 0$ when the convergence is achieved.

Derivatives of the error function (2) with respect to $\boldsymbol{U}$ with $\boldsymbol{V}$ fixed and with respect to $\boldsymbol{V}$ with $\boldsymbol{U}$ fixed, are given by

$$\nabla_U\mathcal{E} = [\nabla_U\mathcal{E}]^+ - [\nabla_U\mathcal{E}]^- = \boldsymbol{U}\boldsymbol{V}^\top\boldsymbol{V} - \boldsymbol{X}\boldsymbol{V}, \tag{6}$$

$$\nabla_V\mathcal{E} = [\nabla_V\mathcal{E}]^+ - [\nabla_V\mathcal{E}]^- = \boldsymbol{V}\boldsymbol{U}^\top\boldsymbol{U} - \boldsymbol{X}^\top\boldsymbol{U}. \tag{7}$$

With these gradient calculations, the rule (5) with $\eta = 1$ yields the well-known Lee and Seung's multiplicative updates [6]

$$\boldsymbol{U} \leftarrow \boldsymbol{U} \odot \frac{\boldsymbol{X}\boldsymbol{V}}{\boldsymbol{U}\boldsymbol{V}^\top\boldsymbol{V}}, \tag{8}$$

$$\boldsymbol{V} \leftarrow \boldsymbol{V} \odot \frac{\boldsymbol{X}^\top\boldsymbol{U}}{\boldsymbol{V}\boldsymbol{U}^\top\boldsymbol{U}}, \tag{9}$$

where the division is also an elementwise operation.

2.2 Probabilistic interpretation and normalization

Probabilistic interpretation of NMF, as in probabilistic latent semantic indexing (PLSI), was given in [7] where equivalence between PLSI and NMF (with I-divergence) was shown.

Let us consider the joint probability of term and document, $p(t_i, d_j)$, which is factorized by

$$\begin{aligned} p(t_i, d_j) &= \sum_k p(t_i, d_j|c_k)p(c_k) \\ &= \sum_k p(t_i|c_k)p(d_j|c_k)p(c_k), \end{aligned} \tag{10}$$

where $p(c_k)$ is the prior probability for cluster c_k. Elements of the term-document matrix, X_{ij}, can be treated as $p(t_i, d_j)$, provided X_{ij} are divided by $\mathbf{1}^\top \boldsymbol{X} \mathbf{1}$ such that $\sum_i \sum_j X_{ij} = 1$ where $\mathbf{1} = [1, \ldots, 1]^\top$ with appropriate dimension.

Relating (10) to the factorization (1), U_{ik} corresponds to $p(t_i|c_k)$, representing the significance of term t_i in cluster c_k. Applying sum-to-one normalization to each column of $\boldsymbol{U}$, i.e., $\boldsymbol{U}\boldsymbol{D}_U^{-1}$ where $\boldsymbol{D}_U \equiv \mathrm{diag}\left(\mathbf{1}^\top \boldsymbol{U}\right)$, we have an exact relation

$$\left[\boldsymbol{U}\boldsymbol{D}_U^{-1}\right]_{ik} = p(t_i|c_k).$$

Assume that $\boldsymbol{X}$ is normalized such that $\sum_i \sum_j X_{ij} = 1$. We define a scaling matrix $\boldsymbol{D}_V \equiv \mathrm{diag}\left(\mathbf{1}^\top \boldsymbol{V}\right)$. Then the factorization (1) can be rewritten as

$$\boldsymbol{X} = (\boldsymbol{U}\boldsymbol{D}_U^{-1})(\boldsymbol{D}_U\boldsymbol{D}_V)(\boldsymbol{V}\boldsymbol{D}_V^{-1})^\top. \tag{11}$$

Comparing (11) with the factorization (10), one can see that each element of the diagonal matrix $\boldsymbol{D} \equiv \boldsymbol{D}_U\boldsymbol{D}_V$ corresponds to cluster prior $p(c_k)$. In the case of unnormalized $\boldsymbol{X}$, the prior matrix $\boldsymbol{D}$ absorb the scaling factor. In practice, the data matrix does not have to be normalized in advance.

In a task of clustering, we need to calculate the posterior of cluster $p(c_k|d_j)$. Applying Bayes' rule, the posterior of cluster is given by the document likelihood and cluster prior probability. That is, $p(c_k|d_j)$ is given by

$$\begin{aligned} p(c_k|d_j) &\propto p(d_j|c_k)p(c_k) \\ &= \left[\boldsymbol{D}(\boldsymbol{V}\boldsymbol{D}_V^{-1})^\top\right]_{kj} \\ &= \left[(\boldsymbol{D}_U\boldsymbol{D}_V)(\boldsymbol{D}_V^{-1}\boldsymbol{V}^\top)\right]_{kj} \\ &= \left[\boldsymbol{D}_U\boldsymbol{V}^\top\right]_{kj}. \end{aligned} \tag{12}$$

It follows from (12) that $(\boldsymbol{V}\boldsymbol{D}_U)^\top$ yields the posterior probability of cluster, requiring the normalization of $\boldsymbol{V}$ using the diagonal matrix $\boldsymbol{D}_U$. Thus, we assign document d_j to cluster k^* if

$$k^* = \arg\max_k [\boldsymbol{V}\boldsymbol{D}_U]_{jk}.$$

Document clustering by NMF was first developed in [4]. Here we use only different normalization and summarize the algorithm below.

Algorithm outline: Document clustering by NMF

---

1. Construct a term-document matrix $\boldsymbol{X}$.
2. Apply NMF to $\boldsymbol{X}$, yielding $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$.
3. Normalize $\boldsymbol{U}$ and $\boldsymbol{V}$:

$$\begin{aligned}\boldsymbol{U} &\leftarrow \boldsymbol{U}\boldsymbol{D}_U^{-1},\\ \boldsymbol{V} &\leftarrow \boldsymbol{V}\boldsymbol{D}_U,\end{aligned}$$

   where $\boldsymbol{D}_U = \boldsymbol{1}^\top\boldsymbol{U}$.
4. Assign document d_j to cluster k^* if

$$k^* = \arg\max_k V_{jk}.$$

---

3 Orthogonal NMF for document clustering

Orthogonal NMF involves a decomposition (1) as in NMF but requires that $\boldsymbol{U}$ or $\boldsymbol{V}$ satisfies the orthogonality constraint such that $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}$ or $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}$ [8]. In this paper we consider the case where $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}$ is incorporated into the optimization (3). In such a case, it was shown that orthogonal NMF is equivalent to k-means clustering in the sense that they share the same objective function [9]. In this section, we present a new algorithm for orthogonal NMF with $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}$ where we incorporate the gradient on Stiefel manifold into multiplicative update.

Orthogonal NMF with $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}$ is formulated as following optimization problem:

$$\begin{aligned}&\arg\min_{U,V} \mathcal{E} = \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^\top\|^2\\ &\text{subject to } \ \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}, \boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0.\end{aligned} \tag{13}$$

In general, the constrained optimization problem (13) is solved by introducing a Lagrangian with a penalty term $\mathrm{tr}\left\{\boldsymbol{\Lambda}(\boldsymbol{V}^\top\boldsymbol{V} - \boldsymbol{I})\right\}$ where $\boldsymbol{\Lambda}$ is a symmetric matrix containing Lagrangian multipliers. Ding *et al.* [3] took this approach with some approximation, developing multiplicative updates.

Here we present a different approach, incorporating the gradient in a constraint surface on which $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}$ is satisfied, into (5). With $\boldsymbol{U}$ fixed in (2), we treat (2) as a function of $\boldsymbol{V}$. Minimizing (2) where $\boldsymbol{V}$ is constrained to the set of $n \times K$ matrices such that $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}$ was well studied in [10, 11]. Here we incorporate nonnegativity constraints on $\boldsymbol{V}$ to develop multiplicative updates with preserving the orthogonality constraint $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}$. The constraint surface

which is the set of $n \times K$ orthonormal matrices such that $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$ is known as the Stiefel manifold [12].

An equation defining tangents to the Stiefel manifold at a point $\boldsymbol{V}$ is obtained by differentiating $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$, yielding

$$\boldsymbol{V}^\top \boldsymbol{\Delta} + \boldsymbol{\Delta}^\top \boldsymbol{V} = 0, \tag{14}$$

i.e., $\boldsymbol{V}^\top \boldsymbol{\Delta}$ is *skew-symmetric*. The canonical metric on the Stiefel manifold [11] is given by

$$g_c(\boldsymbol{\Delta}, \boldsymbol{\Delta}) = \operatorname{tr}\left\{ \boldsymbol{\Delta}^\top \left(\boldsymbol{I} - \frac{1}{2} \boldsymbol{V}\boldsymbol{V}^\top \right) \boldsymbol{\Delta} \right\}, \tag{15}$$

whereas the Euclidean metric is given by

$$g_e(\boldsymbol{\Delta}, \boldsymbol{\Delta}) = \operatorname{tr}\left\{ \boldsymbol{\Delta}^\top \boldsymbol{\Delta} \right\}. \tag{16}$$

We define the partial derivatives of $\mathcal{E}$ with respect to the elements of $\boldsymbol{V}$ as

$$[\nabla_V \mathcal{E}]_{ij} = \frac{\partial \mathcal{E}}{\partial V_{ij}}. \tag{17}$$

For the function $\mathcal{E}$ (2) (with $\boldsymbol{U}$ fixed) defined on the Stiefel manifold, the gradient of $\mathcal{E}$ at $\boldsymbol{V}$ is defined to be the tangent vector $\widetilde{\nabla}_V \mathcal{E}$ such that

$$\begin{aligned} g_e\left(\nabla_V \mathcal{E}, \boldsymbol{\Delta}\right) &= \operatorname{tr}\left\{ \left(\nabla_V \mathcal{E}\right)^\top \boldsymbol{\Delta} \right\} \\ &= g_c\left(\widetilde{\nabla}_V \mathcal{E}, \boldsymbol{\Delta} \right) \\ &= \operatorname{tr}\left\{ \left(\widetilde{\nabla}_V \mathcal{E} \right)^\top \left(\boldsymbol{I} - \frac{1}{2} \boldsymbol{V}\boldsymbol{V}^\top \right) \boldsymbol{\Delta} \right\}, \end{aligned} \tag{18}$$

for all tangent vectors $\boldsymbol{\Delta}$ at $\boldsymbol{V}$.

Solving (18) for $\widetilde{\nabla}_V \mathcal{E}$ such that $\boldsymbol{V}^\top \widetilde{\nabla}_V \mathcal{E}$ is skew-symmetric yields

$$\widetilde{\nabla}_V \mathcal{E} = \nabla_V \mathcal{E} - \boldsymbol{V} (\nabla_V \mathcal{E})^\top \boldsymbol{V}. \tag{19}$$

Thus, with partial derivatives in (7), the gradient in the Stiefel manifold is calculated as

$$\begin{aligned} \widetilde{\nabla}_V \mathcal{E} &= (-\boldsymbol{X}^\top \boldsymbol{U} + \boldsymbol{V}\boldsymbol{U}^\top \boldsymbol{U}) - \boldsymbol{V}(-\boldsymbol{X}^\top \boldsymbol{U} + \boldsymbol{V}\boldsymbol{U}^\top \boldsymbol{U})^\top \boldsymbol{V} \\ &= \boldsymbol{V}\boldsymbol{U}^\top \boldsymbol{X}\boldsymbol{V} - \boldsymbol{X}^\top \boldsymbol{U} \\ &= [\widetilde{\nabla}_V \mathcal{E}]^+ - [\widetilde{\nabla}_V \mathcal{E}]^-. \end{aligned} \tag{20}$$

Invoking the relation (5) with replacing ∇_V by $\widetilde{\nabla}_V$ yields

$$\boldsymbol{V} \leftarrow \boldsymbol{V} \odot \frac{\boldsymbol{X}^\top \boldsymbol{U}}{\boldsymbol{V}\boldsymbol{U}^\top \boldsymbol{X}\boldsymbol{V}}, \tag{21}$$

which is our ONMF algorithm. The updating rule for $\boldsymbol{U}$ is the same as (8).

4 Experiments

We tested our orthogonal NMF algorithm on the six standard document datasets (CSTR, k1a, k1b, re0, and re1) and compared the performance with the standard NMF and the Ding *et al.*'s orthogonal NMF (DTPP)[3]. We applied the stemming and stop-word removal for each dataset, and select 1,000 terms based on the mutual information with the class labels. Normalized-cut weighting [4] is applied to the input data matrix.

We use the accuracy to compare the clustering performance of different algorithms. To compute the accuracy, we first applied Kuhn-Munkres maximal matching algorithm [13] to find the appropriate matching between the clustering result and the target labels. If we denote the true label for the document n to be c_n, and the matched label $\tilde{c}_n$, the accuracy AC can be computed by

$$\mathrm{AC} = \frac{\sum_{n=1}^{N} \delta(c_n, \tilde{c}_n)}{N},$$

where $\delta(x, y) = 1$ for $x = y$ and $\delta(x, y) = 0$ for $x \neq y$. Because the algorithms gave different results depending on the initial conditions, we calculated the mean of 100 runs for different initial conditions. Our orthogonal NMF algorithm gave better performance than the standard NMF and DTPP for the most of the datasets (Table 1).

The orthogonality of the matrix $\boldsymbol{V}$ is also measured by using $\|\boldsymbol{V}^T\boldsymbol{V} - \boldsymbol{I}\|$. The changes of the orthogonality over the iterations are measured and averaged for 100 trials. Our orthogonal NMF algorithm obtained better orthogonality than DTPP for the most of the datasets. The change of orthogonality for the CSTR dataset is shown in Fig. 1 for an example.

	NMF	DTPP	ONMF
cstr	0.7568	**0.7844**	0.7268
wap	0.4744	0.4281	**0.4917**
k1a	0.4773	0.4311	**0.4907**
k1b	0.7896	0.6087	**0.8109**
re0	0.3624	0.3384	**0.3691**
re1	0.4822	0.4452	**0.5090**

Table 1: Mean clustering accuracies (n=100) of standard NMF, Ding *et al.*'s orthogonal NMF (DTPP), and our orthogonal NMF (ONMF) for six document datasets.

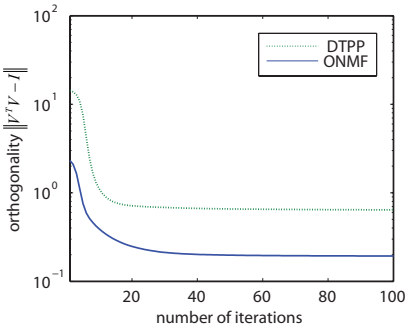

Fig. 1: The orthogonality $\|\boldsymbol{V}^T\boldsymbol{V} - \boldsymbol{I}\|$ convergence of Ding *et al.*'s orthogonal NMF (DTPP) and our orthogonal NMF (ONMF) for the CSTR dataset.

5 Conclusions

We have developed multiplicative updates on Stiefel manifold for orthogonal NMF and have successfully applied it to a task of document clustering, confirming its performance gains over standard NMF and existing orthogonal NMF.

Acknowledgments: This work was supported by National Core Research Center for Systems Bio-Dynamics and Korea Ministry of Knowledge Economy under the ITRC support program supervised by the IITA (IITA-2008-C1090-0801-0045).

References

1. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401** (1999) 788–791
2. Li, S.Z., Hou, X.W., Zhang, H.J., Cheng, Q.S.: Learning spatially localized parts-based representation. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Kauai, Hawaii (2001) 207–212
3. Ding, C., Li, T., Peng, W., Park, H.: Orthogonal nonnegative matrix tri-factorizations for clustering. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Philadelphia,PA (2006)
4. Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In: Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR), Toronto, Canada (2003)
5. Shahnaz, F., Berry, M., Pauca, P., Plemmons, R.: Document clustering using nonnegative matrix factorization. Information Processing and Management **42** (2006) 373–386
6. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems (NIPS). Volume 13., MIT Press (2001)
7. Gaussier, E., Goutte, C.: Relation between PLSA and NMF and implications. In: Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR), Salvador, Brazil (2005)
8. Choi, S.: Algorithms for orthogonal nonnegative matrix factorization. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN), Hong Kong (2008)
9. Ding, C., He, X., Simon, H.D.: On the equivalence of nonnegative matrix factorization and spectral clustering. In: Proceedings of the SIAM International Conference on Data Mining (SDM), Newport Beach, CA (2005) 606–610
10. Smith, S.T.: Geometric Optimization Methods for Adaptive Filtering. PhD thesis, Harvard University (May 1993)
11. Edelman, A., Arias, T., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. **20**(2) (1998) 303–353
12. Stiefel, E.: Richtungsfelder und fernparallelismus in n-dimensionalem mannig faltigkeiten. Commentarii Math. Helvetici **8** (1935-1936) 305–353
13. Lovasz, L., Plummer, M.: Matching Theory. Akademiai Kiado (1986)