2009 International Symposium on Nonlinear Theory and its Applications
NOLTA'09, Sapporo, Japan, October 18-21, 2009

A Derivation of the Trace Difference Optimization Algorithm for the Trace Ratio Optimization Problem

Kohei Inoue, Kenji Hara, and Kiichi Urahama

Faculty of Design, Kyushu University
4-9-1, Shiobaru, Minami-ku, Fukuoka, 815-8540 Japan
Email: {k-inoue,hara,urahama}@design.kyushu-u.ac.jp

Abstract—It is well known that a large number of problems for dimensionality reduction result in the trace ratio optimization problem (TROP). Recently, Wang et al. have proposed an iterative procedure for solving TROP. They transform TROP into a trace difference optimization problem (TDOP) which is efficiently solved with the eigenvalue decomposition method. However, the mechanism of the transformation of TROP into TDOP is not clear in their papers. In this paper, we derive the TDO algorithm for TROP on the basis of the Lagrange multipliers. Moreover, we show that multilinear principal component analysis proposed by Lu et al. recently is a special case of tensor subspace learning which is also formulated as a TROP.

1. Introduction

Dimensionality reduction is one of the most fundamental research topics in computer vision and pattern recognition. The commonly used dimensionality reduction methods include supervised approaches such as linear discriminant analysis (LDA) [1, 2], and unsupervised ones such as principal component analysis (PCA) [3]. These methods are vector-based, i.e., input data are always arranged in a vector form regardless of the original data representations. It has been found that the vector-based methods encounter the singularity problem intrinsically. To address this problem, several extensions of both LDA and PCA have been presented. For matrix-based LDA methods, Ye et al. [4] proposed two-dimensional linear discriminant analysis (2DLDA), and Cai et al. [5] proposed tensor LDA. For higher-order tensor-based LDA method, Yan et al. [6, 7] proposed discriminant analysis with tensor representation (DATER) or multilinear discriminant analysis (MDA). For matrix-based PCA methods, Yang et al. [8] proposed two-dimensional PCA (2DPCA), and Ye et al. [9] proposed generalized PCA (GPCA), and Cai et al. [5] proposed tensor PCA. For higher-order tensor-base PCA method, Lu et al. [10] proposed multilinear PCA (MPCA) recently.

Yan et al. [11] proposed a general framework for dimensionality reduction called graph embedding. They showed that previous methods for dimensionality reduction including PCA, LDA, locality preserving projection (LPP) [12], isometric feature mapping (ISOMAP) [13], locally linear embedding (LLE) [14], and Laplacian eigenmap [15] can be reformulated into the graph embedding framework. Tensor subspace analysis (TSA) proposed by He et al. [16] is also included in the framework [21]. Each method included in the graph embedding framework is formulated as a trace ratio optimization problem (TROP). Conventionally, TROP is often simplified to a ratio trace optimization problem (RTOP) as summarized in [2, 17, 18, 19], since RTOP can be reduced to the corresponding generalized eigenvalue problem which can be solved analytically. Wang et al. [20, 21] proposed an efficient iterative procedure to solve TROP directly, and proved that the value of the trace ratio increases monotonically in the procedure. In each iteration of the procedure, a TROP is transformed into a trace difference optimization problem (TDOP) which is efficiently solved with the eigenvalue decomposition method [2]. Recently, the TDO algorithm is utilized in tensor linear Laplacian discrimination (TLLD) proposed by Zhang et al. [22]. However, the mechanism of the transformation of TROP into TDOP is not clear in their papers [20, 21]. Recently, Nie et al. [23] analysed TROP for vector data and derived a faster TDO algorithm, and proposed semi-supervised orthogonal discriminant analysis.

In this paper, we derive the TDO algorithm proposed by Wang et al. [20, 21] for TROP for higher-order tensor data on the basis of the Lagrange multipliers. Moreover, we show that MPCA proposed by Lu et al. [10] recently is a special case of tensor subspace learning (TSL) [21] which is also included in the graph embedding framework [11].

The rest of this paper is organized as follows. Section 2 summarizes TROP and TDOP for TSL. Section 3 derives TDO algorithm. Section 4 shows a relationship between MPCA and TSL. Section 5 concludes this paper.

2. Trace Ratio Optimization Problem for Tensor Subspace Learning

Let $G = \{\mathcal{X}, S\}$ be an undirected graph with a set of vertices $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_M\}$ and a similarity matrix $S = [s_{mm'}] \in \mathbb{R}^{M\times M}$, where $\mathcal{X}_m = [x_{mi_1\ldots i_N}] \in \mathbb{R}^{I_1\times\cdots\times I_N}$ for $m = 1, \ldots, M$ is an Nth-order tensor where $x_{mi_1\ldots i_N}$ is the $(i_1, \ldots, i_N)$ element of $\mathcal{X}_m$. In graph embedding [11], G is referred to as an intrinsic graph and another graph $H = \{\mathcal{X}, P\}$, where $P = [p_{mm'}] \in \mathbb{R}^{M\times M}$, is also

taken into account and referred to as a penalty graph. Let $U = \{U^{(1)}, \ldots, U^{(N)}\}$ be a set of matrices, where $U^{(n)} = [u^{(n)}_{i_n j_n}] \in \mathbb{R}^{I_n \times J_n}$, $J_n \leq I_n$ for $n = 1, \ldots, N$. Then the sequence of the n-mode products [24] of $\mathcal{X}_m$ and $U^T = \{U^{(1)^T}, \ldots, U^{(N)^T}\}$, where $U^{(n)^T}$ denotes the transpose of $U^{(n)}$, is denoted by $\mathcal{Y}_m = \mathcal{X}_m \times \{U^T\} = \mathcal{X}_m \times_1 U^{(1)^T} \cdots \times_N U^{(N)^T}$ [25], where $\mathcal{X}_m \times_n U^{(n)^T} = [\sum_{i_n=1}^{I_n} x_{m i_1 \ldots i_n \ldots i_N} u^{(n)}_{i_n j_n}] \in \mathbb{R}^{I_1 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \times \cdots \times I_N}$ is the n-mode product [24] of $\mathcal{X}_m$ and $U^{(n)^T}$, and $\mathcal{Y}_m = [y_{m j_1 \ldots j_N}] \in \mathbb{R}^{J_1 \times \cdots \times J_N}$ where $y_{m j_1 \ldots j_N}$ is the $(j_1, \ldots, j_N)$ element of $\mathcal{Y}_m$. Then the trace ratio optimization problem (TROP) for tensor subspace learning (TSL) is expressed as follows:

$$\max_U \quad \frac{\sum_{m=1}^{M} \sum_{m'=1}^{M} \|\mathcal{Y}_m - \mathcal{Y}_{m'}\|_F^2 \, p_{mm'}}{\sum_{m=1}^{M} \sum_{m'=1}^{M} \|\mathcal{Y}_m - \mathcal{Y}_{m'}\|_F^2 \, s_{mm'}}. \tag{1}$$

Let $f(U)$ be the objective funtion in (1). Then we have

$$\begin{aligned} f(U) &= \frac{\sum_{m=1}^{M} \sum_{m'=1}^{M} \left\| U^{(n)^T} \left(X^{(-n)}_{m(n)} - X^{(-n)}_{m'(n)} \right) \right\|_F^2 p_{mm'}}{\sum_{m=1}^{M} \sum_{m'=1}^{M} \left\| U^{(n)^T} \left(X^{(-n)}_{m(n)} - X^{(-n)}_{m'(n)} \right) \right\|_F^2 s_{mm'}} \quad (2) \\ &= \frac{\mathrm{tr}\left(U^{(n)^T} P^{(n)} U^{(n)} \right)}{\mathrm{tr}\left(U^{(n)^T} S^{(n)} U^{(n)} \right)}, \quad (3) \end{aligned}$$

where $X^{(-n)}_{m(n)}$ denotes the monde-n matricizing [25] of $\mathcal{X}_m \times_{-n} \{U^T\} = \mathcal{X}_m \times_1 U^{(1)^T} \cdots \times_{n-1} U^{(n-1)^T} \times_{n+1} U^{(n+1)^T} \cdots \times_N U^{(N)^T}$ [25], and $P^{(n)} = \sum_{m=1}^{M} \sum_{m'=1}^{M} p_{mm'} \left(X^{(-n)}_{m(n)} - X^{(-n)}_{m'(n)} \right) \left(X^{(-n)}_{m(n)} - X^{(-n)}_{m'(n)} \right)^T$, $S^{(n)} = \sum_{m=1}^{M} \sum_{m'=1}^{M} s_{mm'} \left(X^{(-n)}_{m(n)} - X^{(-n)}_{m'(n)} \right) \left(X^{(-n)}_{m(n)} - X^{(-n)}_{m'(n)} \right)^T$. Wang et al. [20, 21] proposed an iterative algorithm for solving TROP by transforming it into a trace difference optimization problem (TDOP) defined for each n by

$$\max_{U^{(n)}} \quad \mathrm{tr}\left[U^{(n)^T} \left(P^{(n)} - \lambda S^{(n)} \right) U^{(n)} \right] \tag{4}$$

subject to the constraint $\left(U^{(n)} \right)^T U^{(n)} = I_{J_n}$, where λ is the value of $f(U)$ into which U obtained in the previous iteration is substituted, and I_{J_n} is the $J_n \times J_n$ identity matrix. The procedure for renewing $U^{(n)}$ for $n = 1, \ldots, N$ is iterated until the convergence. Wang et al. [20, 21] proved that $f(U)$ increases monotonically and the projection matrices U converge. However, the mechanism of the transformation of TROP into TDOP is not clear in their papers [20, 21]. In the next section, we show a relationship between TROP and TDOP on the basis of the Lagrange multipliers.

3. A Derivation of the Trace Difference Optimization Algorithm

In the trace ratio optimization, i.e., the maximization of (3), we may normalize the denominator as $\mathrm{tr}[(U^{(n)})^T S^{(n)} U^{(n)}] = c$ for a constant c while the value of $f(U)$ keeps unchanged by multiplying $\mathrm{tr}[(U^{(n)})^T P^{(n)} U^{(n)}]$ by some appropriate scalar because the numerator and the denominator multiplied by the same scalar result in the same fraction. Therefore, we may reformulate TROP as follows:

$$\max_{U^{(n)}} \quad \mathrm{tr}\left(U^{(n)^T} P^{(n)} U^{(n)} \right) \tag{5}$$

$$\text{subj.to} \quad \mathrm{tr}\left(U^{(n)^T} S^{(n)} U^{(n)} \right) = c, \tag{6}$$

$$U^{(n)^T} U^{(n)} = I_{J_n}. \tag{7}$$

The Lagrange function for this constrained optimization problem is defined by

$$\begin{aligned} L\left(U^{(n)}, \lambda, \Lambda^{(n)} \right) &= \mathrm{tr}\left(U^{(n)^T} P^{(n)} U^{(n)} \right) \\ &\quad - \lambda \left[\mathrm{tr}\left(U^{(n)^T} S^{(n)} U^{(n)} \right) - c \right] \\ &\quad - \mathrm{tr}\left[\Lambda^{(n)} \left(U^{(n)^T} U^{(n)} - I_{J_n} \right) \right], \quad (8) \end{aligned}$$

where λ is a Lagrange multiplier and $\Lambda^{(n)} \in \mathbb{R}^{J_n \times J_n}$ is a symmetric matrix of which the elements are also the Lagrange multipliers. Then we have a necessary condition for optimality as follows:

$$\frac{1}{2} \frac{\partial L}{\partial U^{(n)}} = P^{(n)} U^{(n)} - \lambda S^{(n)} U^{(n)} - U^{(n)} \Lambda^{(n)} = 0, \tag{9}$$

from which we have

$$U^{(n)^T} \left(P^{(n)} - \lambda S^{(n)} \right) U^{(n)} = \Lambda^{(n)}. \tag{10}$$

That is, the solutions of this equation are a diagonal matrix $\Lambda^{(n)}$ of which the diagonal elements are the largest J_n eigenvalues of $P^{(n)} - \lambda S^{(n)}$ and the corresponding eigenvectors which are stacked in $U^{(n)}$, which is also a solution to TDOP (4) and satisfies the orthogonal constraint (7) automatically.

We next consider another optimization problem, i.e., (5) with (6) except (7). The Lagrange function for this constrained optimization problem is defined by

$$\begin{aligned} \tilde{L}\left(U^{(n)}, \lambda \right) &= \mathrm{tr}\left(U^{(n)^T} P^{(n)} U^{(n)} \right) \\ &\quad - \lambda \left[\mathrm{tr}\left(U^{(n)^T} S^{(n)} U^{(n)} \right) - c \right]. \quad (11) \end{aligned}$$

From $\partial \tilde{L} / \partial U^{(n)} = 0$, we have

$$P^{(n)} U^{(n)} = \lambda S^{(n)} U^{(n)}. \tag{12}$$

Multiplying both sides of this equation by $U^{(n)T}$ from left and taking the trace, we have

$$\lambda = \frac{\mathrm{tr}\left(U^{(n)T} P^{(n)} U^{(n)}\right)}{\mathrm{tr}\left(U^{(n)T} S^{(n)} U^{(n)}\right)}, \tag{13}$$

which is identical with $f(U)$.

Consequently, we have derived two equations (10) and (13) for TDOP from two constrained optimization problems, which are the reformulations of TROP, on the basis of the Lagrange multipliers. Both of the two problems attempt to maximize the same objective function $\mathrm{tr}\left(U^{(n)T} P^{(n)} U^{(n)}\right)$ subject to different constraints. Although (13) is derived from a relaxed version of TROP, i.e., from which the orthogonal constraint (7) is excepted in the relaxed TROP, it is expected that the TDO algorithm will converge because $U^{(n)}$ satisfying (7) is always substituted into (13) in the TDO algorithm. In practice, Wang et al. [20, 21] proved the convergency of the algorithm.

4. A Relationship between Tensor Subspace Learning and Multilinear Principal Component Analysis

In this section, we show that multilinear principal component analysis (MPCA) [10] is a special case of tensor subspace learning (TSL) [21].

MPCA is formulated as the maximization of the total scatter of tensors as follows:

$$\max_{U} \quad \Psi_{\mathcal{Y}} \tag{14}$$

$$\text{subj.to} \quad U^{(n)T} U^{(n)} = I_{J_n}, \quad n = 1, \ldots, N, \tag{15}$$

where $\Psi_{\mathcal{Y}}$ is the total scatter defined by $\Psi_{\mathcal{Y}} = \sum_{m=1}^{M} \left\|\mathcal{Y}_m - \bar{\mathcal{Y}}\right\|_F^2$ where $\bar{\mathcal{Y}} = \frac{1}{M}\sum_{m=1}^{M} \mathcal{Y}_m$. On the other hand, TSL can be reformulated as follows:

$$\max_{U} \quad \sum_{m=1}^{M}\sum_{m'=1}^{M} \|\mathcal{Y}_m - \mathcal{Y}_{m'}\|_F^2 \, p_{mm'} \tag{16}$$

$$\text{subj.to} \quad \sum_{m=1}^{M}\sum_{m'=1}^{M} \|\mathcal{Y}_m - \mathcal{Y}_{m'}\|_F^2 \, s_{mm'} = c, \tag{17}$$

$$U^{(n)T} U^{(n)} = I_{J_n}, \quad n = 1, \ldots, N. \tag{18}$$

Let E be the objective function in (16). Assume that

$$p_{mm'} = 1, \quad m, m' = 1, \ldots, M. \tag{19}$$

Then we have

$$\begin{aligned}
E &= \sum_{m=1}^{M}\sum_{m'=1}^{M} \|\mathcal{Y}_m - \mathcal{Y}_{m'}\|_F^2 && (20)\\
&= \sum_{m=1}^{M}\sum_{m'=1}^{M} \left\|Y_{m(n)} - Y_{m'(n)}\right\|_F^2 && (21)\\
&= \sum_{m=1}^{M}\sum_{m'=1}^{M} \mathrm{tr}\left[\left(Y_{m(n)} - Y_{m'(n)}\right)\left(Y_{m(n)} - Y_{m'(n)}\right)^T\right] && (22)\\
&= 2M\sum_{m=1}^{M} \left\|Y_{m(n)}\right\|_F^2 - 2\sum_{m=1}^{M}\sum_{m'=1}^{M} \mathrm{tr}\left(Y_{m(n)} Y_{m'(n)}^T\right) && (23)\\
&= 2M\Bigg[\sum_{m=1}^{M} \left\|Y_{m(n)}\right\|_F^2 - \frac{1}{M}\sum_{m=1}^{M}\sum_{m'=1}^{M} \mathrm{tr}\left(Y_{m(n)} Y_{m'(n)}^T\right)\\
&\quad -\frac{1}{M}\sum_{m=1}^{M}\sum_{m''=1}^{M} \mathrm{tr}\left(Y_{m(n)} Y_{m''(n)}^T\right)\\
&\quad +\frac{1}{M}\sum_{m'=1}^{M}\sum_{m''=1}^{M} \mathrm{tr}\left(Y_{m'(n)} Y_{m''(n)}^T\right)\Bigg] && (24)\\
&= 2M\sum_{m=1}^{M} \mathrm{tr}\Bigg(Y_{m(n)} Y_{m(n)}^T - \frac{1}{M}\sum_{m'=1}^{M} Y_{m(n)} Y_{m'(n)}^T\\
&\quad - \frac{1}{M}\sum_{m''=1}^{M} Y_{m(n)} Y_{m''(n)}^T\\
&\quad +\frac{1}{M^2}\sum_{m'=1}^{M}\sum_{m''=1}^{M} Y_{m'(n)} Y_{m''(n)}^T\Bigg) && (25)\\
&= 2M\sum_{m=1}^{M} \mathrm{tr}\Bigg[\left(Y_{m(n)} - \frac{1}{M}\sum_{m'=1}^{M} Y_{m'(n)}\right)\\
&\quad \left(Y_{m(n)} - \frac{1}{M}\sum_{m''=1}^{M} Y_{m''(n)}\right)^T\Bigg] && (26)\\
&= 2M\sum_{m=1}^{M} \left\|Y_{m(n)} - \frac{1}{M}\sum_{m'=1}^{M} Y_{m'(n)}\right\|_F^2 && (27)\\
&= 2M\sum_{m=1}^{M} \left\|\mathcal{Y}_m - \bar{\mathcal{Y}}\right\|_F^2, && (28)
\end{aligned}$$

where $Y_{m(n)}$ denotes the mode-n matricizing [25] of $\mathcal{Y}_m$. Hence, we have

$$E = 2M\Psi_{\mathcal{Y}}. \tag{29}$$

That is, if (19) is assumed, then the maximization of E is equivalent to that of $\Psi_{\mathcal{Y}}$. For (17), assume that

$$s_{mm'} = \delta_{mm'} \quad m, m' = 1, \ldots, M, \tag{30}$$

where δ is the Kronecker delta. Then we have

$$\sum_{m=1}^{M}\sum_{m'=1}^{M} \|\mathcal{Y}_m - \mathcal{Y}_{m'}\|_F^2 \, s_{mm'} = 0. \tag{31}$$

Thus, the constraint (17) vanishes. Consequently, we may conclude that, if (19) and (30) are assumed, then TSL is reduced to MPCA.

5. Conclusion

In this paper, we have derived the trace difference optimization algorithm for the trace ratio optimization problem on the basis of the Lagrange multipliers. Moreover, we have shown that multilinear principal component analysis is a special case of tensor subspace learning.

Acknowledgments

This work was partially supported by the Ministry of Education, Culture, Sports, Science and Technology under the Grant-in-Aid for Scientific Research (20700165).

References

[1] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[2] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press Professional, 2nd ed., 1990.

[3] I.T. Jolliffe, *Principal Component Analysis*, Springer, 2nd edition, 2002.

[4] J. Ye, R. Janardan, and Qi Li, "Two-dimensional linear discriminant analysis," *Proc. NIPS*, vol. 17, pp. 1569–1576, 2004.

[5] D. Cai, X.F. He, and J.W. Han, "Subspace learning based on tensor analysis," Department of Computer Science Technical Report No. 2572, University of Illinois at Urbana-Champaign (UIUCDCS-R-2005-2572), 2005.

[6] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H.-J. Zhang, "Discriminant analysis with tensor representation," *Proc. CVPR*, vol. 1, pp. 526–532, 2005.

[7] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H.-J. Zhang, "Multilinear discriminant analysis for face recognition," *IEEE Trans. Image Processing*, vol. 16, no. 1, pp. 212–220, 2007.

[8] J. Yang, D. Zhang, A.F. Frangi, and J. Yang, "Two-dimensional PCA: a new approach to appearance-based face representation and recognition," *IEEE Trans. PAMI*, vol. 26, no. 1, pp. 131–137, 2004.

[9] J. Ye, R. Janardan, and Q. Li, "GPCA: an efficient dimension reduction scheme for image compression and retrieval," *Proc. KDD*, pp. 354–363, 2004.

[10] H. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "MPCA: multilinear principal component analysis of tensor objects," *IEEE Trans. Neural Networks*, vol. 19, no. 1, pp. 18–39, 2008.

[11] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Trans. PAMI*, vol. 29, No. 1, pp. 40–51, 2007.

[12] X. He and P. Niyogi, "Locality preserving projections," *Proc. NIPS*, 2003.

[13] J.B. Tenenbaum, V. de Silva, and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[14] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[15] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Proc. NIPS*, 2001.

[16] X. He, D. Cai, and P. Niyogi, "Tensor subspace analysis," *Proc. NIPS*, 2005.

[17] J. Ye and Q. Li, "LDA/QR: an efficient and effective dimension reduction algorithm and its theoretical foundation," *Pattern Recognition*, vol. 37, no. 4, pp. 851–854, 2004.

[18] S. Ji and J. Ye, "A unified framework for generalized linear discriminant analysis," *Proc. CVPR*, 2008.

[19] S. Ji and J. Ye, "Generalized linear discriminant analysis: a unified framework and efficient model selection," *IEEE Trans. Neural Networks*, vol. 19, no. 10, pp. 1768–1782, 2008.

[20] H. Wang, S. Yan, T. Huang, and X. Tang, "Trace ratio vs. ratio trace for dimensionality reduction," *Proc. CVPR*, 2007.

[21] H. Wang, S. Yan, T. Huang, and X. Tang, "A convergent solution to tensor subspace learning," *Proc. IJCAI*, 2007.

[22] W. Zhang, Z. Lin, and X. Tang, "Tensor linear Laplacian discrimination (TLLD) for feature extraction," *Pattern Recognition*, vol. 42, no. 9, pp. 1941–1948, 2009.

[23] F. Nie, S. Xiang, Y. Jia, and C. Zhang, "Semi-supervised orthogonal discriminant analysis via label propagation," *Pattern Recognition*, vol. 42, no. 11, pp. 2615–2627, 2009.

[24] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[25] B.W. Bader and T.G. Kolda, "Algorithm 862: MATLAB tensor classes for fast algorithm prototyping," *ACM Trans. Math. Software*, vol. 32, no. 4, pp. 635–653, 2006.