Bernoulli **17**(4), 2011, 1386–1399
DOI: 10.3150/10-BEJ317

Tensor-based projection depth

YONGGANG HU[*], YONG WANG[**] and YI WU[†]

Department of Mathematics and Systems Science, National University of Defense Technology, Changsha 410073, China. E-mail: [*]xd7688@hotmail.com; [**]yongwang82@gmail.com; [†]wuyi_work@sina.com

The conventional definition of a depth function is vector-based. In this paper, a novel projection depth (PD) technique directly based on tensors, such as matrices, is instead proposed. Tensor projection depth (TPD) is still an ideal depth function and its computation can be achieved through the iteration of PD. Furthermore, we also discuss the cases for sparse samples and higher order tensors. Experimental results in data classification with the two projection depths show that TPD performs much better than PD for data with a natural tensor form, and even when the data have a natural vector form, TPD appears to perform no worse than PD.

Keywords: data depth; Rayleigh projection depth; statistical depth; tensor-based projection depth

1. Introduction

In the last ten years, statistical depth functions have increasingly served as a useful tool in multidimensional exploratory data analysis and inference. The depth of a point in the multidimensional space measures the centrality of that point with respect to a multivariate distribution or a given multivariate data cloud. Depth functions have been successfully used in many fields, such as quality indices [17,20], multivariable regression [24], limiting p values [18], robust estimation [3], nonparametric tests [4] and discriminant analysis [6,11,12,14]. Some common statistical depths which have been defined include *half-space depth* [25], *simplicial depth* [19], *projection depth* [7,8,23,29], *spatial depth* [26], *spatial rank depth* [10] and *integrated dual depth* [5]. Compared to the others, projection depth (PD) is preferable because of its good properties such as robustness, affine invariance, maximality at center, monotonicity relative to deepest point, vanishing at infinity and so on.

However, almost all the depths proposed in the literature are defined over the vector space by now, and the fact is that not all of the observations are naturally in vector form. In the real world, the extracted feature of an object often has some specialized structures, and such structures are in the form of a second, or even higher order tensor. For example, this is the case when a captured image is a second-order tensor, that is, a matrix, and when the sequential data, such as a video sequence for event analysis, is in the form of a third-order tensor. It would be desirable to keep the underlying structures of the data unchanged during the data analysis.

Most of the previous work on depth has first transformed the input tensor data into vectors, which in fact changes the underlying structure of the data sets. At the same time, such a transformation often leads to the curse of dimensionality problem and the small sample size problem since most depth functions (such as Mahalanobis depth) require the covariance matrix to be positive definite.

1350-7265

Therefore, it is necessary to extend the definition of depth to tensor spaces in order to process the data sets directly with tensors without modifying the structures of them. In fact, many tensor-based methods in discriminant analysis have been proposed and have led to many nice results [2,27,28]. In this paper, informed by the aforementioned works, we propose a tensor-based projection depth (TPD) in order to extend the definition of projection depth to tensor spaces. We will prove that TPD is still an ideal depth according to the criteria [30]. Also, we will explore the characteristics of high order tensor projection depth in theory. We will demonstrate that TPD allows us to avoid the above two problems when using vector representation.

The paper is organized as follows. Section 2 briefly introduces tensor algebra. Section 3 introduces the projection depth and gives the solution to the Rayleigh projection depth. Section 4 gives the definition of tensor projection depth and discusses its properties. Section 5 supplies the algorithm for TPD and analyzes its convergence. Section 6 analyzes the special case of sparse samples. Section 7 discusses the TPD for higher order tensors. Section 8 gives numerical results for TPD. Section 9 concludes the paper, and proofs of selected theorems and propositions are given in the Appendix.

2. Tensor algebra

A tensor T of order k is a real-valued multilinear function on k vector spaces [13]:

$$T : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k} \to \mathbb{R}.$$

A multilinear function is linear as a function of each variable considered separately. The set of all kth-order tensors on $\mathbb{R}^{n_i}$, $i = 1, \ldots, k$, denoted by $\mathcal{T}^k$, is a vector space under the usual operations of pointwise addition and scalar multiplication:

$$(aT)(\mathbf{a}_1, \ldots, \mathbf{a}_k) = a(T(\mathbf{a}_1, \ldots, \mathbf{a}_k)),$$
$$(T + T')(\mathbf{a}_1, \ldots, \mathbf{a}_k) = T(\mathbf{a}_1, \ldots, \mathbf{a}_k) + T'(\mathbf{a}_1, \ldots, \mathbf{a}_k),$$

where $\mathbf{a}_i \in \mathbb{R}^{n_i}$.

Given two tensors, $S \in \mathcal{T}^k$ and $T \in \mathcal{T}^l$, their product,

$$S \otimes T : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_{k+l}} \to \mathbb{R},$$

is defined as

$$S \otimes T(\mathbf{a}_1, \ldots, \mathbf{a}_{k+l}) = S(\mathbf{a}_1, \ldots, \mathbf{a}_k) T(\mathbf{a}_{k+1}, \ldots, \mathbf{a}_{k+l}).$$

It is immediate from the multilinearity of S and T that $S \otimes T$ depends linearly on each argument $\mathbf{a}_i$ separately, so it is a $(k + l)$th-order tensor.

First-order tensors are simply vectors on $\mathbb{R}^{n_1}$. That is, $\mathcal{T}_1 = \mathcal{R}^{n_1}$, where $\mathcal{R}^{n_1}$ is the dual space of $\mathbb{R}^{n_1}$. A second-order tensor space is a product of two first-order tensor spaces, that is, $\mathcal{T}^2 = \mathcal{R}^{n_1} \otimes \mathcal{R}^{n_2}$. Let $\mathbf{e}_1, \ldots, \mathbf{e}_{n_1}$ be the standard basis of $\mathbb{R}^{n_1}$ and $\varepsilon_1, \ldots, \varepsilon_{n_1}$ be the dual basis [21] of

$\mathcal{R}^{n_1}$ which is formed from coordinate functions with respect to the basis of $\mathcal{R}^{n_1}$. Likewise, let $\tilde{\mathbf{e}}_1, \ldots, \tilde{\mathbf{e}}_{n_1}$ be a basis of $\mathbb{R}^{n_2}$ and $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_{n_1}$ be the dual basis of $\mathcal{R}^{n_2}$. We have

$$\varepsilon_i(\mathbf{e}_j) = \delta_{ij} \quad \text{and} \quad \tilde{\varepsilon}_i(\tilde{\mathbf{e}}_j) = \delta_{ij},$$

where δ_{ij} is the Kronecker delta function. Thus, $\{\varepsilon_i \otimes \tilde{\varepsilon}_j\}$ $(1 \le i \le n_1, 1 \le j \le n_2)$ forms a basis for $\mathcal{R}^{n_1} \otimes \mathcal{R}^{n_2}$. For any second-order tensor T, we can write

$$T = \sum_{i,j} T_{ij} \varepsilon_i \otimes \tilde{\varepsilon}_j.$$

Given two vectors $\mathbf{a} = \sum_{k=1}^{n_1} a_k \mathbf{e}_k \in \mathbb{R}^{n_1}$ and $\mathbf{b} = \sum_{l=1}^{n_2} b_l \tilde{\mathbf{e}}_l \in \mathbb{R}^{n_2}$, we have

$$\begin{aligned} T(\mathbf{a}, \mathbf{b}) &= \sum_{ij} T_{ij} \varepsilon_i \otimes \tilde{\varepsilon}_j \left(\sum_{k=1}^{n_1} a_k \mathbf{e}_k, \sum_{l=1}^{n_2} b_l \tilde{\mathbf{e}}_l \right) \\ &= \sum_{ij} T_{ij} \varepsilon_i \left(\sum_{k=1}^{n_1} a_k \mathbf{e}_k \right) \tilde{\varepsilon}_j \left(\sum_{l=1}^{n_2} b_l \tilde{\mathbf{e}}_l \right) \\ &= \sum_{ij} T_{ij} a_i b_j = \mathbf{a}^T T \mathbf{b}. \end{aligned} \tag{2.1}$$

This shows that every second-order tensor in $\mathcal{R}^{n_1} \otimes \mathcal{R}^{n_2}$ uniquely corresponds to an $n_1 \times n_2$ matrix.

Note that in this paper, our primary interest is focused on second-order tensors. However, most of our conclusions for second-order TPD can be naturally extended to higher orders. We will discuss this question in Section 7.

3. Projection depth

According to [29], the definition of projection depth can be expressed as follows.

Definition 3.1. *Let μ and σ be univariate location and scale measures, respectively. Define the outlyingness of a point $\mathbf{x} \in \mathbb{R}^p$ with respect to a given function F of X in $\mathbb{R}^p$, $p \ge 1$, as*

$$O(\mathbf{x}, F) = \sup_{\|\mathbf{u}\|=1} \frac{|\mathbf{u}^T \mathbf{x} - \mu(F_{\mathbf{u}})|}{\sigma(F_{\mathbf{u}})}, \tag{3.1}$$

where $F_{\mathbf{u}}$ is the distribution of $\mathbf{u}^T X$. Then, $O(\mathbf{x}, F)$ is defined to be 0 if $\mathbf{u}^T \mathbf{x} - \mu(F_{\mathbf{u}}) = \sigma(F_{\mathbf{u}}) = 0$. The projection depth (PD) of a point $\mathbf{x} \in \mathbb{R}^p$ with respect to the given F, $PD(\mathbf{x}, F)$, is then defined as

$$PD(\mathbf{x}, F) = \frac{1}{1 + O(\mathbf{x}, F)}. \tag{3.2}$$

Remark 3.1. Here, we also assume that μ and σ exist uniquely, μ is translation and scale equivariant, and σ is scale equivariant and translation invariant, that is, $\mu(F_{sY+c}) = s\mu(F_Y) + c$ and $\sigma(F_{sY+c}) = |s|\sigma(F_Y)$, respectively, for any scalars s, c and random variable $Y \in \mathbb{R}^1$.

The most popular outlying function is defined as

$$O(\mathbf{x}, F) = \sup_{\|\mathbf{u}\|=1} \frac{|\mathbf{u}^T\mathbf{x} - Med(F_{\mathbf{u}})|}{MAD(F_{\mathbf{u}})}, \tag{3.3}$$

where $F_{\mathbf{u}}$ is the distribution of $\mathbf{u}^T X$, $Med(F_{\mathbf{u}})$ is the median of $F_{\mathbf{u}}$ and $MAD(F_{\mathbf{u}})$ is the median of the distribution of $|\mathbf{u}^T X - Med(F_u)|$.

Apart from the good properties of a statistical depth function, this version of PD is more robust compared with other depths. However, it is hard to compute for high-dimensional samples.

Obviously, the variance and mean are also natural choices for σ and μ, respectively. It is easy to prove that such a projection-based depth is also an ideal depth function. And, most importantly, its computation is very simple.

Theorem 3.1 (Rayleigh projection depth). *Let* $(\mu, \sigma) =$ (mean, variance), *and suppose that the second moments of* X *exist and that* $X \sim F$. *The solution of the outlying function* (3.1) *is then that of a Rayleigh quotient problem*,

$$\begin{aligned} O_R(\mathbf{x}, F) &= \sup_{\|\mathbf{u}\|=1} \frac{|\mathbf{u}^T\mathbf{x} - E(\mathbf{u}^T X)|}{\sqrt{E(\mathbf{u}^T X - E(\mathbf{u}^T X))^2}} \\ &= \sqrt{\frac{\mathbf{u}_1^T \mathbf{A}\mathbf{u}_1}{\mathbf{u}_1^T \mathbf{B}\mathbf{u}_1}} = \sqrt{\lambda_1}, \end{aligned} \tag{3.4}$$

where $\mathbf{A}$ *is the matrix* $(\mathbf{x} - EX)(\mathbf{x} - EX)^T$, $\mathbf{B}$ *is* $E(X - EX)(X - EX)^T$, λ_1 *is the largest eigenvalue of the generalized eigenvalue problem*

$$\mathbf{A}\mathbf{z} = \lambda\mathbf{B}\mathbf{z}, \qquad \mathbf{z} \neq 0,$$

and $\mathbf{u}_1$ *is the corresponding eigenvector of* λ_1.

We call this projection depth the Rayleigh projection depth.

Remark 3.2. In this paper, for the convenience of computation, the examples in the experiments are all based on the Rayleigh projection depth, that is, $(\mu, \sigma) =$ (mean, varance).

Remark 3.3. Obviously, RPD requires the covariance $\mathbf{B}$ to be positive. To avoid this situation, for the sparse samples, we simply project the samples into their nonzero subspace using principal component analysis (PCA).

4. Tensor projection depth

Before describing tensor projection depth, we first review the terminology associated with tensor operations [15,16]. The inner product of tensors $\mathbf{A}$ and $\mathbf{B}$ (with the same orders and dimensions) is $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i,j} \mathbf{A}_{ij}\mathbf{B}_{ij}$. The norm of a tensor $\mathbf{A}$ is defined as its Frobenius norm, that is, $\|\mathbf{A}\| = \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle}$, and the distance between two tensors $\mathbf{A}$ and $\mathbf{B}$ in $\mathcal{R}^{n_1} \otimes \mathcal{R}^{n_2}$ is defined as $\|\mathbf{A} - \mathbf{B}\|$, where $\mathbf{A} - \mathbf{B} = (\mathbf{A}_{ij} - \mathbf{B}_{ij})_{n_1 \times n_2}$.

From the tensorial viewpoint, if we take X as a random variable in the first-order tensor space $\mathcal{R}^{n_1}$, then the outlyingness of the projection depth in Definition 3.1 can be expressed as

$$O(\mathbf{x}, X) = \sup_{\|\mathbf{u}\|=1} \frac{|\mathbf{x}(\mathbf{u}) - \mu(X(\mathbf{u}))|}{\sigma(X(\mathbf{u}))}.$$

Thus, if $\mathcal{X} \in \mathcal{R}^{n_1} \otimes \mathcal{R}^{n_2}$ is a random variable, then, according to the formula (2.1), the outlying function in the tensor space $\mathcal{R}^{n_1} \otimes \mathcal{R}^{n_2}$ can be naturally defined as

$$O(\mathbf{X}, \mathcal{X}) = \sup_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} \frac{|\mathbf{X}(\mathbf{u}, \mathbf{v}) - \mu(\mathcal{X}(\mathbf{u}, \mathbf{v}))|}{\sigma(\mathcal{X}(\mathbf{u}, \mathbf{v}))} = \sup_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} \frac{|\mathbf{u}^T\mathbf{X}\mathbf{v} - \mu(\mathbf{u}^T\mathcal{X}\mathbf{v})|}{\sigma(\mathbf{u}^T\mathcal{X}\mathbf{v})}, \tag{4.1}$$

where $\mathbf{u} \in \mathbb{R}^{n_1}$ and $\mathbf{v} \in \mathbb{R}^{n_2}$.

Definition 4.1 (Tensor projection depth). *The projection depth with outlying function given by formula (4.1) is called tensor projection depth.*

For a given univariate location (or "center") measure μ, a distribution function $F_{\mathcal{X}}$ is called *μ-symmetric about the point $\theta \in \mathcal{R}^{n_1} \otimes \mathcal{R}^{n_2}$* if $\mu(\mathbf{u}^T\mathcal{X}\mathbf{v}) = \mathbf{u}^T\theta\mathbf{v}$ for any pair of unit vectors $\mathbf{u} \in \mathbb{R}^{n_1}$, $\mathbf{v} \in \mathbb{R}^{n_2}$. We have the following theorem.

Theorem 4.1. *Suppose that θ in $\mathcal{R}^{n_1} \otimes \mathcal{R}^{n_2}$ is the point of symmetry of a distribution $F(\mathcal{X})$ with respect to a given notion of symmetry. The tensor projection depth function $TPD(\mathbf{X}, \mathcal{X})$ is*:

1. *convex*;
2. *symmetric for μ-symmetric F*;
3. *affine invariant*;
4. *monotonic relative to the deepest point*;
5. *vanishing at infinity, that is, $TPD(\mathbf{X}, \mathcal{X}) \to 0$ as $\|\mathbf{X}\| \to \infty$*;
6. *maximized at the center of μ-symmetric F.*

Remark 4.1. Theorem 4.1 shows that TPD is still an ideal depth according to the criteria [30]. Furthermore, we can easily obtain many other properties of TPD beyond those of the PD in [29], such as the properties of its sample versions and its medians. However, these are not the key points of this paper and so we omit any detailed discussion here.

5. Algorithm

Suppose that the elements of $S_n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ are generated from F (where F_n is its empirical distribution) and that $\mathbf{X}$ is a fixed tensor. The TPD of $\mathbf{X}$ with respect to F_n can then be computed by the following algorithm:

1. *Initialization:* Let $\mathbf{u} = (1, \ldots, 1)^T$.
2. *Computing* $\mathbf{v}$: Let $\mathbf{x}_i = \mathbf{X}_i^T \mathbf{u}$ and $F_n^{\mathbf{u}} = \mathbf{u}^T F_n$. Then, $\mathbf{v}$ can be computed by solving the vector-based projection depth

$$\sup_{\|\mathbf{v}\|=1} \frac{|\mathbf{v}^T \mathbf{x} - \mu(F_n^{\mathbf{u}} \mathbf{v})|}{\sigma(F_n^{\mathbf{u}} \mathbf{v})}. \tag{5.1}$$

3. *Computing* $\mathbf{u}$*:* Once $\mathbf{v}$ is obtained, let $\tilde{\mathbf{x}}_i = \mathbf{X}_i \mathbf{v}$ and $F_n^{\mathbf{v}} = F_n \mathbf{v}$. Then, $\mathbf{u}$ can be computed by solving the following optimization problem:

$$\sup_{\|\mathbf{u}\|=1} \frac{|\mathbf{u}^T \tilde{\mathbf{x}} - \mu(\mathbf{u}^T F_n^{\mathbf{v}})|}{\sigma(\mathbf{u}^T F_n^{\mathbf{v}})}. \tag{5.2}$$

4. *Iteratively computing* $\mathbf{u}$ *and* $\mathbf{v}$: Using steps 2 and 3, we can iteratively compute $\mathbf{u}$ and $\mathbf{v}$ until they tend to converge.

Remark 5.1. The optimization problems (5.1) and (5.2) are the same as (3.3) in the vector-based projection depth algorithm. Thus, any computational method for the projection depth can also be used here.

The following theorem shows that the above algorithm converges.

Theorem 5.1. *The iterative procedure to solve the optimization problems* (5.1) *and* (5.2) *will monotonically increase the objective function value in* (4.1)*, hence the algorithm converges.*

Remark 5.2. Furthermore, if the optimization problem (3.1) is convex, then the solution of (4.1) is also globally optimal. For instance, if $(\mu, \sigma) = (\text{mean, variance})$ (the Rayleigh projection depth), then its solution is also globally optimal.

6. Sparse samples

As with RPD, TPD based on RPD also faces the problem of sparse samples. From formulas (5.1) and (5.2), we know that for any sample set $S_n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and its corresponding empirical distribution F_n, the algorithm in the previous section requires the covariance matrices of $F_n^{\mathbf{u}}$ and $F_n^{\mathbf{v}}$ to be positive for any $\mathbf{u} \in \mathbb{R}^{n_1}$ and $\mathbf{v} \in \mathbb{R}^{n_2}$. However, in practice, the tensor data usually do not satisfy such requirements.

There are two factors that can lead to such non-positiveness. First, the sample size is too small, that is, the size of S_n is less than n_1 or n_2. Second, the data have some common columns or rows

(e.g., the images have identical color edges or patterns). In the vector space, we usually use PCA to remove the redundant null space of the samples and therefore we can use the tensor PCA proposed by Cai *et al.* [1] to reduce the dimensionality of the tensor samples.

Suppose that $M_{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$,

$$M_V = \sum_{i=1}^{n} \big((\mathbf{X}_i - M_{\mathbf{X}})(\mathbf{X}_i - M_{\mathbf{X}})^T\big),$$

$$M_U = \sum_{i=1}^{n} \big((\mathbf{X}_i - M_{\mathbf{X}})^T(\mathbf{X}_i - M_{\mathbf{X}})\big),$$

where the columns of V are the eigenvectors of M_V, and U are the eigenvectors of M_U. Thus, the new mappings of F_n can be expressed as

$$F_n^{(r_1,r_2)} = \{V_{r_1}^T \mathbf{X}_1 U_{r_2}, \ldots, V_{r_1}^T \mathbf{X}_n U_{r_2}\}, \tag{6.1}$$

where r_1 and r_2 are the mapping dimensions, and V_{r_1} and U_{r_2} are the first r_1 and r_2 columns of V and U, respectively. Here, we take r_1 and r_2 to be the ranks of M_V and M_U.

Theorem 6.1. *For any* $\mathbf{u} \in \mathbb{R}^{r_1}$, $\mathbf{v} \in \mathbb{R}^{r_2}$ *with* $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$, *the covariance matrices of* $\mathbf{u}^T F_n^{(r_1,r_2)}$ *and* $F_n^{(r_1,r_2)}\mathbf{v}$ *are always positive.*

7. Higher order tensors

The algorithm described above takes second-order tensors (i.e., matrices) as input data. However, the algorithm can also be extended to higher order tensors. In this section, we briefly describe the TPD algorithm for higher order tensors.

Let $S_n = \{\mathbf{X}_i, i = 1, \ldots, n\}$ denote the sample set and F_n its empirical distribution, where $\mathbf{X}_i \in \mathcal{R}^{n_1} \otimes \cdots \otimes \mathcal{R}^{n_k}$. The outlying function of TPD is then

$$O(\mathbf{X}, \mathcal{X}) \doteq \sup_{\|\mathbf{u}_1\|=\cdots=\|\mathbf{u}_k\|=1} \frac{|\mathbf{X}(\mathbf{u}_1, \ldots, \mathbf{u}_k) - \mu(F_n(\mathbf{u}_1, \ldots, \mathbf{u}_k))|}{\sigma(F_n(\mathbf{u}_1, \ldots, \mathbf{u}_k))}, \tag{7.1}$$

where $\mathbf{u}_i \in \mathbb{R}^{n_i}$.

Before stating the algorithm, we first introduce an item of notation which we will need. If $T \in \mathcal{R}^{n_1} \otimes \cdots \otimes \mathcal{R}^{n_k}$, then for any $\mathbf{a}_l \in \mathbb{R}^{n_l}$, $1 \le l \le k$, we use $T \times_l \mathbf{a}_l$ to denote a new tensor in $\mathcal{R}^{n_1} \otimes \cdots \otimes \mathcal{R}^{n_{l-1}} \otimes \mathcal{R}^{n_{l+1}} \otimes \cdots \otimes \mathcal{R}^{n_k}$, namely

$$T \times_l \mathbf{a}_l = \sum_{i_l=1}^{n_l} T_{i_1,\ldots,i_{l-1},\ldots,i_{l+1},\ldots,i_k} \cdot a_{i_l}. \tag{7.2}$$

Thus, the algorithm for higher order tensors can naturally be expressed as follows:

1. *Initialization*: Let $\mathbf{u}_i^0 = (x_1, \ldots, x_{n_i})^T$, $x_j \in \mathbb{R}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, k-1$.

2. *Computing* $\mathbf{u}_k^0$: If we let $\mathbf{x}^k = \mathbf{X} \times_1 \mathbf{u}_1^0 \times_2 \mathbf{u}_2^0 \times \cdots \times_{k-1} \mathbf{u}_{k-1}^0$, then $\mathbf{u}_k^0$ can be computed by solving the vector-based projection depth

$$\sup_{\|\mathbf{u}_k^0\|=1} \frac{|\mathbf{u}_k^{0T}\mathbf{x}^k - \mu(F_n \times_1 \mathbf{u}_1^0 \times_2 \mathbf{u}_2^0 \times \cdots \times_{k-1} \mathbf{u}_{k-1}^0)|}{\sigma(F_n \times_1 \mathbf{u}_1^0 \times_2 \mathbf{u}_2^0 \times \cdots \times_{k-1} \mathbf{u}_{k-1}^0)}. \tag{7.3}$$

3. *Computing* $\mathbf{u}_{k-1}^1$: Once $\mathbf{u}_k^0$ is obtained, we let $\mathbf{x}^{k-1} = \mathbf{X} \times_1 \mathbf{u}_1^0 \times \cdots \times_{k-2} \mathbf{u}_{k-2}^0 \times_k \mathbf{u}_k^0$ and $\mathbf{u}_{k-1}^1$ can be computed by solving the optimization problem

$$\sup_{\|\mathbf{u}_{k-1}^1\|=1} \frac{|\mathbf{u}_{k-1}^{1T}\mathbf{x}^{k-1} - \mu(F_n \times_1 \mathbf{u}_1^0 \times \cdots \times_{k-2} \mathbf{u}_{k-2}^0 \times_k \mathbf{u}_k^0)|}{\sigma(F_n \times_1 \mathbf{u}_1^0 \times \cdots \times_{k-2} \mathbf{u}_{k-2}^0 \times_k \mathbf{u}_k^0)}. \tag{7.4}$$

4. *Iteratively computing* $\mathbf{u}_i$, $i = 1, \ldots, k$, until they tend to converge.

Remark 7.1. It is easy to prove that TPD in a higher order tensor space still satisfies the above theorems and that its convergence is also guaranteed by Theorem 5.1.

8. Experiments

First, we use data classification to demonstrate the validity of the TPD. Consider a multivariate data set C that is partitioned into given classes $C_1, \ldots, C_q$. An additional data point $\mathbf{x}$ has to be assigned to one of several given classes of *object*. Suppose that there are q classes. The most natural classifier provided by [14] is then

$$classd(\mathbf{x}) = \arg\max_j D(\mathbf{x}|C_j), \tag{8.1}$$

where $D(\mathbf{x}|C_j)$ is the depth of the $\mathbf{x}$ with respect to class C_j, $i = 1, \ldots, q$. This assigns $\mathbf{x}$ to the class C_j in which $\mathbf{x}$ is deepest.

The Columbia Object Image Library (COIL-20) [22] is a database of grayscale images of 20 objects. The objects were placed on a motorized turntable against a black background. The turntable was rotated through 360 degrees to vary the object pose with respect to a fixed camera. Images of the objects were taken at pose intervals of 5 degrees. This corresponds to 72 images with the dimensions of 32×32 pixels per object. Here, we only take the first 10 objects as examples.

In the experiments, recognition rates under different training sizes are computed by means of the following steps:

1. *Select the test sets*: Randomly select p test sets $\mathbf{X}_{\text{test}}^j$ from the object set $\mathbf{X}_j$ for each class, where $j = 1, \ldots, 10$.
2. *for* each training size n_k
   for each repeating round t:

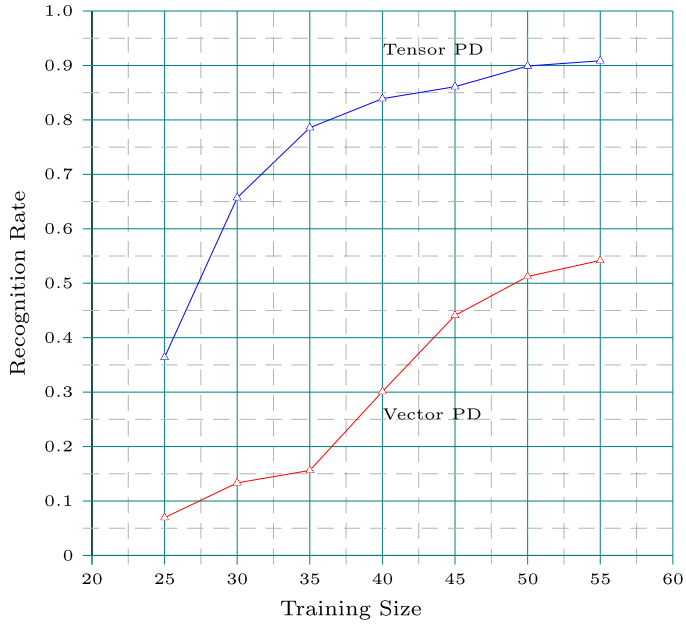

Figure 1. Recognition rates of TPD and PD under different training sample sizes.

- *Randomly select the training sets*: Randomly select n_k training sets from $\mathbf{X}_j/\mathbf{X}_{\text{test}}^j$ (the left samples of $\mathbf{X}_j$) for each j, $j = 1, \ldots, 10$.
- *Compute the recognition rate*. Compute the correctly recognized number ℓ_j for each test set $\mathbf{X}_{\text{test}}^j$ by using the formula (8.1) and compute the glossary recognition rate by $\eta_t = \sum_{j=1}^{10} \ell_j/10p$.

3. Compute the mean and variance of η_t.

Here, $p = 7$ and the training number equals 25, 30, 35, 40, 45, 50, 55, respectively. The results are shown in Figure 1 and Table 1.

From Figure 1 and Table 1, we can see that for such samples with intrinsic tensor form, TPD performs better than PD. A question then naturally arises: If the data sets are naturally in vector form, how does TPD perform compared with PD? We will answer the question by means of the following experiment.

We consider the famous Iris data [9], which contains measurements of four different features (sepal length, sepal width, petal length and petal width) for each of 150 observations from three different types of iris plant: (1) setosa; (2) virginica; (3) versicolor. We randomly choose 10 observations from each class to construct the test sets and then randomly select 10, 15, 20, 25, 30, 35, 40 samples from the remaining observations as the respective training sets. For the computation of TPD, the samples are reshaped as 2×2.

Table 1. The mean, deviation and variance of the recognition rates by TPD and PD with the COIL-20 set

Training size	Tensor projection depth				Projection depth			
	Mean	Min.	Max.	Variance	Mean	Min.	Max.	Variance
25	0.3638	0.2571	0.4143	0.0018	0.0695	0.0286	0.1571	0.0017
30	0.6571	0.5286	0.7429	0.0028	0.1333	0.0571	0.2429	0.0028
35	0.7857	0.7000	0.8571	0.0017	0.1526	0.0429	0.2571	0.0044
40	0.8390	0.7714	0.9000	0.0010	0.3010	0.1714	0.3714	0.0029
45	0.8610	0.7714	0.9429	0.0019	0.4410	0.3571	0.5571	0.0041
50	0.8990	0.8429	0.9286	0.0008	0.5124	0.4429	0.6000	0.0017
55	0.9086	0.8714	0.9571	0.0006	0.5419	0.4429	0.6286	0.0021

From Table 2 we can see that there is no apparent difference between the two results. Therefore, data from vector spaces can be converted into tensors and we can perform the depth procession with TPD.

9. Discussion and conclusion

In this paper, tensor projection depth is proposed as an extension of the definition of depth to tensor spaces. We show that, according to the criteria [30], TPD satisfies all four desirable properties. TPD has the advantages of avoiding the curse of dimensionality and keeping the natural structures of the data sets invariant. For sparse samples, we use tensor PCA to remove their null space and compute the TPD in the subspace. The numerical results show that TPD performs better than PD for data which are naturally in tensor form.

Data sets which are naturally in vector form can also be processed using TPD, which converts the data into tensor form. Although such processing will actually change the structure of the data sets to some extent, numerical results show that there are no apparent differences in the out-

Table 2. The mean, deviation and variance of the recognition rates by TPD and PD with the Iris set

Training size	Tensor projection depth				Projection depth			
	Mean	Min.	Max.	Variance	Mean	Min.	Max.	Variance
10	0.9698	0.8571	1.0000	0.0016	0.9476	0.7619	1.0000	0.0041
15	0.9889	0.9524	1.0000	0.0004	0.9841	0.9524	1.0000	0.0005
20	0.9952	0.9048	1.0000	0.0004	0.9921	0.8571	1.0000	0.0008
25	0.9984	0.9524	1.0000	0.0001	0.9984	0.9524	1.0000	0.0001
30	0.9984	0.9524	1.0000	0.0001	0.9984	0.9524	1.0000	0.0001
35	1.0000	1.0000	1.0000	0.0000	1.0000	1.0000	1.0000	0.0000
40	1.0000	1.0000	1.0000	0.0000	1.0000	1.0000	1.0000	0.0000

come. For some (μ, σ), such tensor-based processing can effectively decrease the computational complexity of PD caused by the dimensionality.

Appendix: Proofs

Proof of Theorem 4.1. *Convexity*. We will show that the outlying function (4.1) is still convex. Let $\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{R}^{n_1} \otimes \mathcal{R}^{n_2}$ be two arbitrary points, $0 < \lambda < 1$, and for the point $\mathbf{X}_0 \doteq (1-\lambda)\mathbf{X}_1 + \lambda\mathbf{X}_2$, we have

$$\begin{aligned}
&|\mathbf{u}^T\mathbf{X}_0\mathbf{v} - \mu(\mathbf{u}^T\mathcal{X}\mathbf{v})| \\
&\quad = \left|(1-\lambda)\big(\mathbf{u}^T\mathbf{X}_1\mathbf{v} - \mu(\mathbf{u}^T\mathcal{X}\mathbf{v})\big) + \lambda\big(\mathbf{u}^T\mathbf{X}_2\mathbf{v} - \mu(\mathbf{u}^T\mathcal{X}\mathbf{v})\big)\right| \\
&\quad \le (1-\lambda)|\mathbf{u}^T\mathbf{X}_1\mathbf{v} - \mu(\mathbf{u}^T\mathcal{X}\mathbf{v})| + \lambda|\mathbf{u}^T\mathbf{X}_2\mathbf{v} - \mu(\mathbf{u}^T\mathcal{X}\mathbf{v})|
\end{aligned}$$

and

$$\begin{aligned}
O(\mathbf{X}_0, \mathcal{X}) &= \sup_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} \frac{|\mathbf{u}^T\mathbf{X}_0\mathbf{v} - \mu(\mathbf{u}^T\mathcal{X}\mathbf{v})|}{\sigma(\mathbf{u}^T\mathcal{X}\mathbf{v})} \\
&\le \sup_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} \frac{(1-\lambda)|\mathbf{u}^T\mathbf{X}_1\mathbf{v} - \mu(\mathbf{u}^T\mathcal{X}\mathbf{v})| + \lambda|\mathbf{u}^T\mathbf{X}_2\mathbf{v} - \mu(\mathbf{u}^T\mathcal{X}\mathbf{v})|}{\sigma(\mathbf{u}^T\mathcal{X}\mathbf{v})} \\
&= (1-\lambda)O(\mathbf{X}_1, \mathcal{X}) + \lambda O(\mathbf{X}_2, \mathcal{X}).
\end{aligned}$$

Thus,

$$TPD(\mathbf{X}_0, \mathcal{X}) \ge (1-\lambda)TPD(\mathbf{X}_1, \mathcal{X}) + \lambda TPD(\mathbf{X}_2, \mathcal{X}).$$

Symmetry. This is straightforward.

Affine invariance. Suppose that $\mathbf{A}_{n_1\times n_1}$ and $\mathbf{B}_{n_2\times n_2}$ are any two non-singular matrices. We then have

$$O(\mathbf{AXB}, \mathbf{A}\mathcal{X}\mathbf{B}) = \sup_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} \frac{|\mathbf{u}^T\mathbf{AXBv} - \mu(\mathbf{u}^T\mathbf{A}\mathcal{X}\mathbf{Bv})|}{\sigma(\mathbf{u}^T\mathbf{A}\mathcal{X}\mathbf{Bv})}.$$

For fixed $\mathbf{X}$, suppose that

$$(\mathbf{u}_0, \mathbf{v}_0) = \arg \sup_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} O(\mathbf{X}, \mathcal{X}).$$

Thus, if we fix $\mathbf{v} = \mathbf{v}_0$ and let

$$\mathbf{u}_1 = \arg \sup_{\|\mathbf{u}\|=1} \frac{|\mathbf{u}^T\mathbf{A}(\mathbf{X}\mathbf{v}_0) - \mu(\mathbf{u}^T\mathbf{A}(\mathcal{X}\mathbf{v}_0))|}{\sigma(\mathbf{u}^T\mathbf{A}(\mathcal{X}\mathbf{v}_0))},$$

then, according to Theorem 2.1 in [29],

$$\sup_{\|\mathbf{u}\|=1} \frac{|\mathbf{u}^T\mathbf{A}(\mathbf{X}\mathbf{v}_0) - \mu(\mathbf{u}^T\mathbf{A}(\mathcal{X}\mathbf{v}_0))|}{\sigma(\mathbf{u}^T\mathbf{A}(\mathcal{X}\mathbf{v}_0))} = \sup_{\|\mathbf{u}\|=1} \frac{|\mathbf{u}^T(\mathbf{X}\mathbf{v}_0) - \mu(\mathbf{u}^T(\mathcal{X}\mathbf{v}_0))|}{\sigma(\mathbf{u}^T(\mathcal{X}\mathbf{v}_0))}.$$

Thus, $\mathbf{u}_1\mathbf{A} = \lambda\mathbf{u}_0$, where $\lambda \in \mathbb{R}$, and we have

$$\begin{aligned}\sup_{\|\mathbf{v}\|=1} \frac{|\mathbf{u}_1^T\mathbf{A}(\mathbf{X}\mathbf{v}) - \mu(\mathbf{u}_1^T\mathbf{A}(\mathcal{X}\mathbf{v}))|}{\sigma(\mathbf{u}_1^T\mathbf{A}(\mathcal{X}\mathbf{v}))} &= \sup_{\|\mathbf{v}\|=1} \frac{|\lambda\mathbf{u}_0^T(\mathbf{X}\mathbf{v}) - \mu(\lambda\mathbf{u}_0^T(\mathcal{X}\mathbf{v}_0))|}{\sigma(\lambda\mathbf{u}_0^T(\mathcal{X}\mathbf{v}))} \\ &= \sup_{\|\mathbf{v}\|=1} \frac{|\mathbf{u}_0^T(\mathbf{X}\mathbf{v}) - \mu(\mathbf{u}_0^T(\mathcal{X}\mathbf{v}_0))|}{\sigma(\mathbf{u}_0^T(\mathcal{X}\mathbf{v}))}.\end{aligned}$$

Therefore,

$$\mathbf{v}_1 = \arg\sup_{\|\mathbf{v}\|=1} \frac{|\mathbf{u}_1^T\mathbf{A}(\mathbf{X}\mathbf{v}) - \mu(\mathbf{u}_1^T\mathbf{A}(\mathcal{X}\mathbf{v}))|}{\sigma(\mathbf{u}_1^T\mathbf{A}(\mathcal{X}\mathbf{v}))} = \mathbf{v}_0.$$

Similarly,

$$\begin{aligned}\sup_{\|\mathbf{v}\|=1} \frac{|\mathbf{u}_1^T\mathbf{A}\mathbf{X}\mathbf{v} - \mu(\mathbf{u}_1^T\mathbf{A}\mathcal{X}\mathbf{v})|}{\sigma(\mathbf{u}_1^T\mathbf{A}\mathcal{X}\mathbf{v})} &= \sup_{\|\mathbf{v}\|=1} \frac{|\mathbf{u}_1^T\mathbf{A}\mathbf{X}\mathbf{B}\mathbf{v} - \mu(\mathbf{u}_1^T\mathbf{A}\mathcal{X}\mathbf{B}\mathbf{v})|}{\sigma(\mathbf{u}_1^T\mathbf{A}\mathcal{X}\mathbf{B}\mathbf{v})} \\ &= \sup_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} \frac{|\mathbf{u}^T\mathbf{A}\mathbf{X}\mathbf{B}\mathbf{v} - \mu(\mathbf{u}^T\mathbf{A}\mathcal{X}\mathbf{B}\mathbf{v})|}{\sigma(\mathbf{u}^T\mathbf{A}\mathcal{X}\mathbf{B}\mathbf{v})} \\ &= \frac{|\mathbf{u}_0^T\mathbf{A}\mathbf{X}\mathbf{B}\mathbf{v}_0 - \mu(\mathbf{u}_0^T\mathbf{A}\mathcal{X}\mathbf{B}\mathbf{v}_0)|}{\sigma(\mathbf{u}_0^T\mathbf{A}\mathcal{X}\mathbf{B}\mathbf{v}_0)}.\end{aligned}$$

The result then follows.

Monotonicity relative to deepest point. Suppose that $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_c \in \mathcal{R}^{n_1} \otimes \mathcal{R}^{n_2}$, $\mathbf{X}_c$ is the deepest tensor and $\mathbf{X}_1 = \lambda\mathbf{X}_2 + (1-\lambda)\mathbf{X}_c$, $\lambda \in [0,1]$. Then, since

$$O(\mathbf{X}_1, \mathcal{X}) \le (1-\lambda)O(\mathbf{X}_2, \mathcal{X}) + \lambda O(\mathbf{X}_c, \mathcal{X}),$$

we have

$$O(\mathbf{X}_1, \mathcal{X}) - \lambda O(\mathbf{X}_c, \mathcal{X}) \le (1-\lambda)O(\mathbf{X}_1, \mathcal{X}) \le (1-\lambda)O(\mathbf{X}_2, \mathcal{X}).$$

Thus, $O(\mathbf{X}_1, \mathcal{X}) \le O(\mathbf{X}_2, \mathcal{X})$ and $TPD(\mathbf{X}_1, \mathcal{X}) \ge TPD(\mathbf{X}_2, \mathcal{X})$, that is, the tensor projection depth decreases monotonically along any ray emanating from the deepest point.

Maximality at center. Suppose that F is θ-symmetric about a unique point $\mathbf{X}_c \in R^{n_1} \times R^{n_2}$. Then, for any pair of unit vectors $\mathbf{u}, \mathbf{v}$, we have $\mu(\mathbf{u}^T\mathcal{X}\mathbf{v}) = \mathbf{u}^T\mathbf{X}_c\mathbf{v}$ and the result follows.

Vanishing at infinity. This is straightforward. □

Proof of Theorem 5.1. Define

$$f(\mathbf{u},\mathbf{v}) = \sup_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} \frac{|\mathbf{u}^T\mathbf{X}\mathbf{v} - \mu(F_n(\mathbf{u},\mathbf{v}))|}{\sigma(F_n(\mathbf{u},\mathbf{v}))}.$$

Let $\mathbf{u}_0$ be the initial value. Fixing $\mathbf{u}_0$, we get $\mathbf{v}_0$ by solving the optimizations (5.1) and (5.2).

Likewise, fixing $\mathbf{v}_0$, we get $\mathbf{u}_1$ by solving the optimization problem (5.2). Thus, we have

$$f(\mathbf{u}_0,\mathbf{v}_0) \le f(\mathbf{u}_1,\mathbf{v}_0).$$

Finally, we get

$$f(\mathbf{u}_0,\mathbf{v}_0) \le f(\mathbf{u}_1,\mathbf{v}_0) \le f(\mathbf{u}_1,\mathbf{v}_1) \le f(\mathbf{u}_2,\mathbf{v}_1)\cdots.$$

Since f is bounded, it converges. □

Acknowledgements

This work was supported by Grants from the Natural Science Fund of China (Nos 60975038 and 60974124). We thank the two referees for their careful reading and useful comments.

References

[1] Cai, D., He, X. and Han, J. (2005). Subspace learning based on tensor analysis. Technical Report UIUCDCS-R-2005-2572, Dept. Computer Science, Univ. Illinois at Urbana-Champaign.
[2] Cai, D., He, X., Wen, J.R., Han, J. and Ma, W.Y. (2006). Support tensor machines for text categorization. Technical report, Dept. Computer Science, Univ. Illinois at Urbana-Champaign.
[3] Chen, Z. and Tyler, D.E. (2004). On the finite sample breakdown points of redescending m-estimates of location. *Statist. Probab. Lett.* **69** 233–242. MR2089000
[4] Chenouri, S. (2004). Multivariate robust nonparametric inference based on data depth. *Univ. Waterloo* **90** 67–89.
[5] Cuevas, A. and Fraiman, R. (2009). On depth measures and dual statistics. A methodology for dealing with general data. *J. Multivariate Anal.* **100** 753–766. MR2478196
[6] Cui, X., Lin, L. and Yang, G.R. (2008). An extended projection data depth and its applications to discrimination. *Commun. Statist. Theory Methods* **37** 2276–2290. MR2526679
[7] Donoho, D.L. (1982). Breakdown properties of multivariate location estimators. Ph.D. thesis, Dept. Statistics, Harvard Univ.
[8] Donoho, D.L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.* **20** 1803–1827. MR1193313
[9] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Ann. of Eugenics* **7** 179–188.
[10] Gao, Y. (2003). Data depth based on spatial rank. *Statist. Probab. Lett.* **65** 217–225. MR2018033
[11] Ghosh, A.K. and Chaudhuri, P. (2005). On data depth and distribution-free discriminant analysis using separating surfaces. *Bernoulli* **11** 1–27. MR2121452
[12] Ghosh, A.K. and Chaudhuri, P. (2005). On maximum depth and related classifiers. *Board of the Foundation of the Scandinavian of Statistics* **32** 327–350. MR2188677

[13] Itskov, M. (2007). *Tensor Algebra and Tensor Analysis for Engineers—With Applications to Continuum Mechanics*. New York: Springer.
[14] Jörnsten, R. (2004). Clustering and classification based on the l_1 data depth. *J. Multivariate Anal.* **90** 67–89. MR2064937
[15] Kolda, T. (2001). Orthogonal tensor decompositions. *SIAM J. Matrix Anal. Appl.* **23** 243–255. MR1856608
[16] Lathauwer, L., Moor, B. and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* **21** 1253–1278. MR1780272
[17] Liu, R. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *J. Amer. Statist. Assoc.* **88** 252–260. MR1212489
[18] Liu, R. and Singh, K. (1997). Notions of limiting p values based on data depth and bootstrap. *J. Amer. Statist. Assoc.* **92** 266–277. MR1436115
[19] Liu, R.Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18** 405–414. MR1041400
[20] Liu, R.Y. (1992). Data depth and multivariate rank tests. In *L_1-Statistics and Related Methods* (Y. Dodge, ed.) 279–294. Amsterdam: North-Holland. MR1214839
[21] Meise, R. and Vogt, D. (1997). *Introduction to Functional Analysis*, 1st ed. *Oxford Graduate Texts in Mathematics* **2**. Oxford: Clarendon Press and Oxford. MR1483073
[22] Nene, S.A., Nayar, S.K. and Murase, H. (1996). Columbia object image library (coil-20). Technical report, Columbia Univ. Available at http://www.cs.columbia.edu/CAVE/.
[23] Stahel, W.A. (1981). Robuste schatzungen: Infintesimale optimalitat and schatzungen von kovarianzmatrizen. Ph.D. thesis, Zurich.
[24] Tian, X., Vardi, Y. and Zhang, C. (2002). l_1-depth, depth realtive to a model, and robust regression. In *Statistical Data Analysis Based on the L_1-Norm and Related Methods* (Y. Dodge, ed.) 285–299. Basel: Birkhäuser.
[25] Tukey, J.W. (1975). Mathematics and picturing of data. In *Proceedings of the International Congress on Mathematics* 523–531. Montreal: Canad. Math. Congress. MR0426989
[26] Vardi, Y. and Zhang, C. (2000). The multivariate l_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA* **97** 1423–1426. MR1740461
[27] Xu, A.B., Jin, X., Jiang, Y.G. and Guo, P. (2006). Complete two-dimensional PCA for face recognition. In *18th International Conference on Pattern Recognition* **3** 481–484. Hong Kong.
[28] Zhi, R.C. and Ruan, Q.Q. (2008). Facial expression recognition based on two-dimensional discriminant locality preserving projections. *Neurocomputing* **71** 1730–1734.
[29] Zuo, Y. (2003). Projection-based depth functions and associated medians. *Ann. Statist.* **31** 1460–1490. MR2012822
[30] Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Ann. Statist.* **28** 461–482. MR1790005

Received May 2009 and revised June 2010