On the Bounding Boxes Obtained by Principal Component Analysis

Darko Dimitrov* Christian Knauer* Klaus Kriegel* Günter Rote*

Abstract

Principle component analysis (PCA) is a commonly used to compute a bounding box of a point set in $\mathbb{R}^d$. In this paper we give bounds on the approximation factor of PCA bounding boxes of convex polygons in $\mathbb{R}^2$ (lower and upper bounds) and convex polyhedra in $\mathbb{R}^3$ (lower bound).

1 Introduction

Substituting sets of points or complex geometric shapes with their bounding boxes is motivated with many applications. For example, in computer graphics, it is used to maintain hierarchical data structures for fast rendering of a scene or for collision detection. Additional applications include those in shape analysis and shape simplification, or in statistics, for storing and performing range-search queries on a large database of samples.

Computing a minimum-area bounding box of a set of n points in $\mathbb{R}^2$ can be done in $O(n \log n)$ time, for example with the rotating caliper algorithm [9]. O'Rourke [6] presented a deterministic algorithm, a rotating caliper variant in $\mathbb{R}^3$, for computing the exact minimum-volume bounding box of a set of n points in $\mathbb{R}^3$. His algorithm requires $O(n^3)$ time and $O(n)$ space. Barequet and Har-Peled [2] have contributed two $(1+\epsilon)$-approximation algorithms for computing the minimum-volume bounding box problem for point sets in $\mathbb{R}^3$, both with nearly linear complexity. The running times of their algorithms are $O(n + 1/\epsilon^{4.5})$ and $O(n \log n + n/\epsilon^3)$.

Numerous heuristics have been proposed for computing a box which encloses a given set of points. The simplest heuristic is naturally to compute the axis-aligned bounding box of the point set. Two-dimensional variants of this heuristic include the well-known $R-tree$, the *packed* $R-tree$ [7], the R^*-tree [8], the R^+-tree [3], etc. A frequently used heuristic for computing a bounding box of a set of points is based on *principal component analysis*. The principal components of the point set define the axes of the bounding box, and the dimension of the bounding box along an axis is given by the extreme values of the projection of the points on the corresponding axis. Two distinguished applications of this heuristic are OBB-tree [4] and BOXTREE [1], hierarchical bounding box structures, which support efficient collision detection and ray tracing. Computing a bounding box of a set of points in $\mathbb{R}^2$ and $\mathbb{R}^3$ by PCA is quite fast, it requires linear time. To avoid the influence of the distribution of the point set on the directions of the PCs, a possible approach is to consider only the boundary of the convex hull of the point set. Thus, the complexity of the algorithm increases to $O(n \log n)$. The popularity of this heuristic besides its speed, lies in its easy implementation and in the fact that usually, PCA bounding boxes are tight-fitting.

We are not aware of any previous published results about the quality of the bounding boxes obtained by PCA. Here we give guarantees on the approximation factor of bounding boxes of convex polygons in $\mathbb{R}^2$ and convex polyhedra in $\mathbb{R}^3$.

The paper is organized as follows. In section 2 we review the basics of principal component analysis. In Section 3 we give lower bounds on the approximation factor of PCA bounding boxes in $\mathbb{R}^2$ and $\mathbb{R}^3$, and in Section 4 an upper bound in $\mathbb{R}^2$. We conclude with future work and open problems in Section 5.

2 Principal Component Analysis

The central idea and motivation of PCA [5] (also known as the Karhunen-Loeve transform, or the Hotelling transform) is to reduce the dimensionality of a data set by identifying *the most significant directions (principal components)*. Let $X = \{x_1, x_2, \ldots, x_m\}$, where x_i is a d-dimensional vector, and $c = (c_1, c_2, \ldots, c_d)$ be the center of gravity of X. For $1 \leq k \leq d$, we use x_{ik} to denote the kth coordinate of the vector x_i. Given two vectors u and v, we use $\langle u, v \rangle$ to denote their inner product. For any unit vector $v \in \mathbb{R}^d$, the *variance of X in direction v* is

$$var(X, v) = \frac{1}{m} \sum_{i=1}^{m} \langle x_i - c, v \rangle^2. \qquad (1)$$

The most significant direction corresponds to the unit vector v_1 such that $var(X, v_1)$ is maximum. In general, after identifying the j most significant directions $B_j = \{v_1, v_2, \ldots, v_j\}$, the $(j+1)$th most significant direction corresponds to the unit vector v_{j+1} such that $var(X, v_{j+1})$ is maximum among all unit vectors perpendicular to $v_1, v_2, \ldots, v_j$.

*Institut für Informatik, Freie Universität Berlin, Germany, {darko, knauer, kriegel, rote}@inf.fu-berlin.de

It can be verified that for any unit vector $v \in \mathbb{R}^d$,

$$var(X, v) = \langle Cv, v \rangle, \tag{2}$$

where C is the *covariance matrix* of X. C is a symmetric $d \times d$ matrix where the ij-th component, $C_{ij}, 1 \leq i, j \leq d$, is defined as

$$C_{ij} = \frac{1}{m} \sum_{k=1}^{m} (x_{ik} - c_i)(x_{jk} - c_j). \tag{3}$$

The procedure of finding the most significant directions, in the sense mentioned above, can be formulated as an eigenvalue problem. Namely, it can be shown that, if $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ are the eigenvalues of C, then the unit eigenvector v_j for λ_j is the jth most significant direction. It follows that all λ_js are non-negative as $\lambda_j = var(X, v_j)$. Since the matrix C is symmetric positive definite, its eigenvectors are orthogonal. The following result summarizes the above background knowledge on PCA. For any set S of orthogonal unit vectors in $\mathbb{R}^d$, we use $var(X, S)$ to denote $\sum_{v \in S} var(X, v)$.

Lemma 1 *For $1 \leq j \leq d$, let λ_j be the j-th largest eigenvalue of C and let v_j denote the unit eigenvector for λ_j. Let $B_j = \{v_1, v_2, \ldots, v_j\}$, $sp(B_j)$ be the linear subspace spanned by B_j, and $sp(B_j)^\perp$ be the orthogonal complement of $sp(B_j)$. Then $\lambda_1 = \max\{var(X, v) : \text{unit vector } v \text{ in } \mathbb{R}^d\}$ and for any $2 \leq j \leq d$,*

i) $\lambda_j = \max\{var(X, v) : \text{unit vector } v \text{ in } sp(B_{j-1})^\perp\}$.

ii) $\lambda_j = \min\{var(X, v) : \text{unit vector } v \text{ in } sp(B_j)\}$.

iii) $var(X, B_j) \geq var(X, S)$ *for any set S of j orthogonal unit vectors.*

Since the covariance matrix depends on the distribution of the points, there is not necessarily a strong correlation between the eigenvectors and the directions of the axes of the minimal bounding box. Consider for example the situation when a significant number of points is located in a small part of the space, see figure 1. Moreover, by adding points inside or on the boundary of the convex hull of the point set, the PCA bounding box can arbitrarily vary between the minimum-volume bounding box and maximum-volume bounding box of the convex hull of the point set. To overcome this problem, one possible approach is to consider only the points on the boundary of the convex hull of the point set when the covariance matrix is computed. This is the approach we take in the rest of the paper. This lead us to so-called continuous PCA. In that case, X is a continuous set of d-dimensional vectors and it can be verified that the derivations and the lemma above also hold.

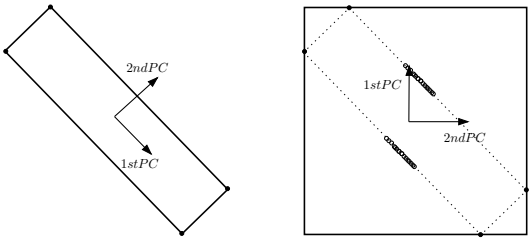

Figure 1: Four points and its PCA bounding-box (left). Dense collection of additional points significantly affect the orientation of the PCA bounding-box (right).

3 Lower Bounds

The following connection between hyperplane reflective symmetry and principal components will help us to derive the lower bounds of the approximation factor of the PCA bounding boxes.

Theorem 2 *Let P be a d-dimensional point set symmetric with respect to a hyperplane H. Then, a principal component of P is orthogonal to H.*

Proof. Without loss of generality, we can assume that the hyperplane of symmetry is spanned by the last $n-1$ standard base vectors of the d-dimensional space and the center of gravity of the point set coincides with the origin of the d-dimensional space, i.e., $c = (0, 0, \ldots, 0)$. Then, the components C_{1j} and C_{j1}, for $2 \leq j \leq d$, are 0 and the covariance matrix has the form:

$$C = \begin{bmatrix} C_{11} & 0 & \ldots & 0 \\ 0 & C_{22} & \ldots & C_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & C_{d2} & \ldots & C_{dd} \end{bmatrix} \tag{4}$$

Its characteristic polynomial has the form:

$$det(C - \lambda\, I) = (C_{11} - \lambda) f(\lambda) \tag{5}$$

where $f(\lambda)$ is a polynomial of degree $d-1$, with coefficients determined by the elements of the $(d-1) \times (d-1)$ submatrix of C. From this it follows that C_{11} is a solution of the characteristic equation, i.e., it is an eigenvalue of C and the vector (1, 0, ...,0) is its corresponding eigenvector (principal component), which is orthogonal to the assumed hyperplane of symmetry. □

3.1 $\mathbb{R}^2$

We obtain a lower bound in $\mathbb{R}^2$ from a rhomb. Let its side length be a. Since the rhomb is symmetric, its PCs coincide with its diagonals. On the left side in figure 2 its optimal-area bounding boxes, for 2 different angles, are shown, and on the right side its

corresponding PCA bounding boxes. As the rhomb's angles approach 90°, its optimal-area bounding box approaches a square with side length a, and the PCA bounding box a square with side length $\sqrt{2}a$. So, the ratio between the area of the PCA bounding box and the area of the optimal-area bounding box in the limit goes to 2.

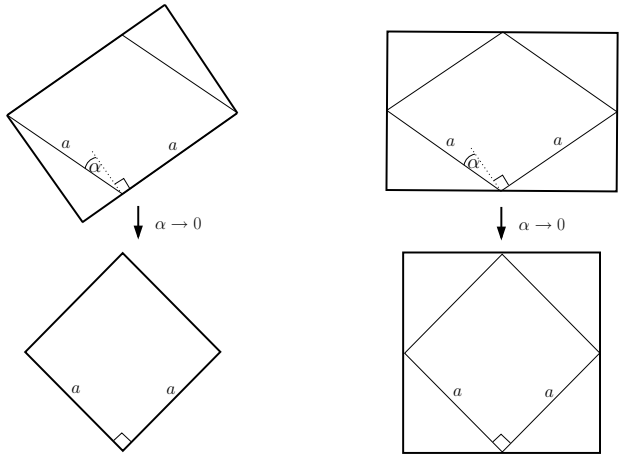

Figure 2: An example which gives us the lower bound of the area of the PCA bounding box of an arbitrary convex polygon in $\mathbb{R}^2$.

Proposition 3 *In general, the ratio between the area of the PCA bounding box and optimal-area bounding box of a convex polygon cannot be smaller than 2.*

3.2 $\mathbb{R}^3$

We obtain a lower bound in $\mathbb{R}^3$ from a square dipyramid, having a rhomb with side length $\sqrt{2}$ as a base. Its other side lengths are $\frac{\sqrt{3}}{2}$. Similarly as in $\mathbb{R}^2$, we consider the case when its base, the rhomb, in limit approaches the square. Then the ratio of the volume of the bounding box on the left side in figure 3, and the volume of its PCA bounding box, on the right in figure 3, goes to 4.

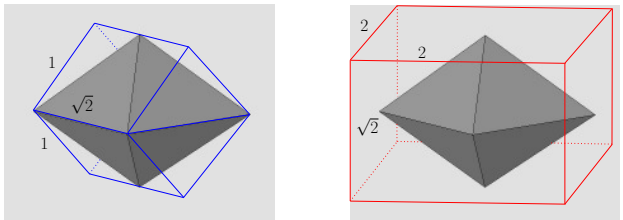

Figure 3: An example which gives the lower bound of the volume of the PCA bounding box of an arbitrary convex polygon in $\mathbb{R}^3$.

Proposition 4 *In general, the ratio between the volume of the PCA bounding box and optimal-volume bounding box of a convex polyhedron cannot be smaller than 4.*

4 Upper Bound in $\mathbb{R}^2$

Let P be a set of points in $\mathbb{R}^2$, and $\mathcal{P}$ the boundary of its convex hull. $\mathcal{P}$, its PCA bounding box and the line l_{pca}, which coincides with the 1st PC of $\mathcal{P}$, are given in the left part of figure 4. The optimal bounding box and the line $l_{\frac{1}{2}}$, going through the middle of its smaller side, parallel with its longer side, are given in the right part of figure 4.

The sides of any bounding box of P, $BB(P)$ (let us denote them with a and b, s.t. $a \geq b$) cannot be larger than the diameter of P. From the other side, it is true that $diam(P) \leq diam(BB(P)) \leq \sqrt{2}a$. So we have the following relation

$$a_{pca} \leq diam(P) \leq \sqrt{2}a_{opt}. \tag{6}$$

We denote with $d^2(\mathcal{P}, l)$ the integral of the squared

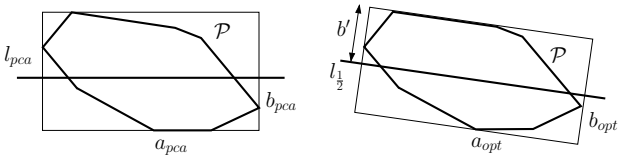

Figure 4: PCA bounding-box and the optimal bounding-box of the polygon $\mathcal{P}$.

distances of the points of $\mathcal{P}$ to the arbitrary line l, i.e. $d^2(\mathcal{P}, l) = \int_{x \in \mathcal{P}} d^2(x, l) ds$. From continuous version of lemma 1, part ii), follows that l_{pca} is the best fitting line in the sense that it minimize the sum of squared distances, and therefore

$$d^2(\mathcal{P}, l_{pca}) \leq d^2(\mathcal{P}, l_{\frac{1}{2}}). \tag{7}$$

We denote with $\mathcal{BB}_{\mathcal{OPT}}$ the boundary of the optimal bounding box of the $\mathcal{P}$. It is true that

$$\begin{aligned} d^2(\mathcal{P}, l_{\frac{1}{2}}) &\leq d^2(\mathcal{BB}_{\mathcal{OPT}}, l_{\frac{1}{2}}) \\ &= \frac{{b_{opt}}^2 a_{opt}}{2} + \frac{{b_{opt}}^3}{6}. \end{aligned} \tag{8}$$

Due to space limitation, we leave the proof of (8) to a full paper. Now we look at $\mathcal{P}$ and its PCA bounding

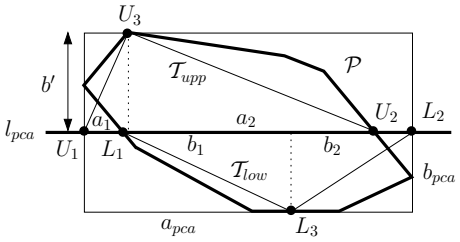

Figure 5: Lower bound for $d^2(\mathcal{P}, l_{pca})$.

box (figure 5). l_{pca} divides $\mathcal{P}$ into an upper and a

lower part, $\mathcal{P}_{upp}$ and $\mathcal{P}_{low}$. Let us denote with l_{upp} the orthogonal projection of $\mathcal{P}_{upp}$ onto l_{pca}, with U_1 and U_2 as its extreme points, and with l_{low} the orthogonal projection of $\mathcal{P}_{low}$ onto l_{pca}, with L_1 and L_2 as its extreme points. Since $\mathcal{P}$ is convex, the following relations hold:

$$|l_{upp}| \geq \frac{b'}{b_{pca}} a_{pca}, \text{ and } |l_{low}| \geq \frac{b_{pca} - b'}{b_{pca}} a_{pca}. \quad (9)$$

We inscribe in $\mathcal{P}_{upp}$ a triangle $\mathcal{T}_{upp}(\triangle U_1U_2U_3)$, and in $\mathcal{P}_{low}$ a triangle $\mathcal{T}_{low}(\triangle L_1L_2L_3)$. It then holds that

$$d^2(\mathcal{P}, l_{pca}) = d^2(\mathcal{P}_{upp} \bigcup \mathcal{P}_{low}, l_{pca}) \geq d^2(\mathcal{T}_{upp} \bigcup \mathcal{T}_{low}, l_{pca}) = d^2(\mathcal{T}_{upp}, l_{pca}) + d^2(\mathcal{T}_{low}, l_{pca}). \quad (10)$$

Due to space limitation, we leave also the proof of (10) to a full paper. The value

$$d^2(\mathcal{T}_{upp}, l_{pca}) = \frac{b'^2}{3}(\sqrt{a_1^2 + b'^2} + \sqrt{a_2^2 + b'^2})$$

is minimal when $a_1 = a_2 = \frac{|l_{upp}|}{2}$. So with (9) we get

$$d^2(\mathcal{T}_{upp}, l_{pca}) \geq \frac{b'^3}{3b_{pca}}(\sqrt{a_{pca}^2 + 4b_{pca}^2}).$$

Analogously, we have for the lower part:

$$d^2(\mathcal{T}_{low}, l_{pca}) \geq \frac{(b_{pca} - b')^3}{3b_{pca}}(\sqrt{a_{pca}^2 + 4b_{pca}^2}).$$

The sum $d^2(\mathcal{T}_{upp}, l_{pca}) + d^2(\mathcal{T}_{low}, l_{pca})$ is minimal when $b' = \frac{b_{pca}}{2}$. This, together with (10), gives:

$$d^2(\mathcal{P}, l_{pca}) \geq \frac{b_{pca}^2}{12}\sqrt{a_{pca}^2 + 4b_{pca}^2}. \quad (11)$$

Combining (7), (8) and (11) we have:

$$\frac{1}{2}a_{opt}b_{opt}^2 + \frac{1}{6}b_{opt}^3 \geq \frac{b_{pca}^2}{12}\sqrt{a_{pca}^2 + 4b_{pca}^2}. \quad (12)$$

Let $a_{pca} = \alpha a_{opt}$ and $b_{pca} = \beta b_{opt}$. Replacing b_{pca}^2 with $\beta^2 b_{opt}^2$ in (12), we obtain:

$$6a_{opt} + 2b_{opt} \geq \beta^2\sqrt{a_{pca}^2 + 4b_{pca}^2} \geq \beta^2 a_{pca}.$$

Replacing further a_{pca} with αa_{opt}, we obtain:

$$6a_{opt} + 2b_{opt} \geq \beta^2\alpha a_{opt} = \frac{\beta^2\alpha}{8}(6a_{opt} + 2a_{opt}).$$

Since $a_{opt} \geq b_{opt}$, we have

$$6a_{opt} + 2b_{opt} \geq \frac{\beta^2\alpha}{8}(6a_{opt} + 2b_{opt}),$$

and from this

$$\beta \leq \sqrt{\frac{8}{\alpha}}. \quad (13)$$

Finally, from (6) and (13) we obtain:

$$\frac{area(BB_{PCA}(\mathcal{P}))}{area(BB_{OPT}(\mathcal{P}))} = \frac{a_{pca}b_{pca}}{a_{opt}b_{opt}} = \alpha\beta \leq \sqrt{8\sqrt{\alpha}} \leq \sqrt{8\sqrt{2}} \approx 3.3635856.$$

We summarize this result in the following theorem:

Theorem 5 *Let $\mathcal{P}$ be the boundary of the convex hull of the point set $P \subset \mathbb{R}^2$. The ratio between the area of the PCA bounding box of $\mathcal{P}$ and the area of its optimal bounding box is bounded from above by* 3.3636.

5 Future Work and Open Problems

Improving the upper bound in $\mathbb{R}^2$, as well as obtaining an upper bound in $\mathbb{R}^3$ are our current interests. A variant of the PCA bounding box problem, where instead of considering only the points on the boundary of the convex hull all points from the convex hull are taken into account, is also of interest. A very demanding open problem is to get an approximation factor of PCA bounding boxes in arbitrary dimension.

References

[1] G. Barequet, B. Chazelle, L. J. Guibas, J. S. B. Mitchell and A. Tal. BOXTREE: A Hierarchical Representation for Surfaces in 3D. *Computer Graphics Forum*, 1996, vol. 15, no.3, pages 387–396.

[2] G. Barequet and S. Har-Peled. Efficiently Approximating the Minimum-Volume Bounding Box of a Point Set in 3D. *Proc. 10th ACM-SIAM Sympos. Discrete Algorithms*, 1999, pages 82–91.

[3] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger. The R*-tree: An efficient and robust access method for points and rectangles. *In ACM SIGMOD Int. Conf. on Manag. of Data*, 1990, pages 322–331.

[4] S. Gottschalk, M. C. Lin and D. Manocha. OBBTree: A Hierarchical Structure for Rapid Interference Detection. *Proc. SIGGRAPH 1996*, pages 171–180.

[5] I. Jolliffe. Principal Component Analysis. *Springer-Verlag, New York, 2nd ed.*, 2002.

[6] J. O'Rourke. Finding Minimal Enclosing Boxes. *Int. J. Comp. Info. Sci. 14 (1985)*, pages 183–199.

[7] N. Roussopoulos and D. Leifker. Direct Spatial Search on Pictorial Databases Using Packed R-Trees. *In Proc. of the ACM SIGMOD*, 1985, pages 17–31.

[8] T. Sellis, N. Roussopoulos, and C. Faloutsos. The R+-tree: A dynamic index for multidimensional objects. *In Proc. 13th VLDB Conference*, 1987, pages 507–518.

[9] G. Toussaint. Solving geometric problems with the rotating calipers. *Proc. IEEE MELECON'83*, May 1983.