Bounds on the Quality of the PCA Bounding Boxes

Darko Dimitrov[a], Christian Knauer[a], Klaus Kriegel[a] and Günter Rote[a]

[a]Institut für Informatik
Freie Universität Berlin
Takustraße 9
D-14195 Berlin, Germany
{darko, knauer, kriegel, rote}@inf.fu-berlin.de

Principal component analysis (PCA) is commonly used to compute a bounding box of a point set in $\mathbb{R}^d$. The popularity of this heuristic lies in its speed, easy implementation and in the fact that usually, PCA bounding boxes quite well approximate the minimum-volume bounding boxes. We present examples of discrete points sets in the plane, showing that the worst case ratio of the volume of the PCA bounding box and the volume of the minimum-volume bounding box tends to infinity. Thus, we concentrate our attention on PCA bounding boxes for continuous sets, especially for the convex hull of a point set. Here, we contribute lower bounds on the approximation factor of PCA bounding boxes of convex sets in arbitrary dimension, and upper bounds in $\mathbb{R}^2$ and $\mathbb{R}^3$.

1. Introduction

Substituting sets of points or complex geometric shapes with their bounding boxes is motivated by 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 calipers algorithm [15]. O'Rourke [12] presented a deterministic algorithm, a rotating calipers variant in $\mathbb{R}^3$, for computing the 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 the minimum-volume bounding box of 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 $O(n \log n + n/\epsilon^3)$, respectively.

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* [13], the *R^*-tree* [3], the *R^+-tree* [14], 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. Once the axis directions are given, the dimension of the bounding box is easily found by the extreme values of the projection of the points on the corresponding axis. Two distinguished applications of this heuristic are the OBB-tree [6] and the 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 simple and 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 the convex hull, or the boundary of the convex hull $CH(P)$ of the point set P. 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, see [10] for some experimental results.

Given a point set $P \subseteq \mathbb{R}^d$ we denote by $BB_{pca}(P)$ the PCA bounding box of P and by $BB_{opt}(P)$ the bounding box of P with smallest possible volume. The ratio of the two volumes $\lambda_d(P) = \mathrm{Vol}(BB_{pca}(P))/\mathrm{Vol}(BB_{opt}(P))$ defines the approximation factor for P, and

$$\lambda_d = \sup\left\{\lambda_d(P) \mid P \subseteq \mathbb{R}^d, \mathrm{Vol}(CH(P)) > 0\right\}$$

defines the general PCA approximation factor. Here, we give lower bounds on λ_d for arbitrary dimension d, and upper bounds on λ_2 and λ_3. [1]

The organization and the main results of the paper are as follows: In Section 2 we review the basics of principal component analysis. In particular, we introduce the continuous version of PCA, which results in a series of approximation factors $\lambda_{d,i}$, where i ranges from 0 to d and denotes the dimension of the faces of the convex hull that contribute to the continuous point set for which the principal components are computed. In Section 3 we give lower bounds on $\lambda_{d,i}$ for arbitrary values of d and $1 \leq i \leq d$. First, we show that $\lambda_{d,i} = \infty$ for any $d \geq 4$ and any $1 \leq i < d-1$. Next, we show that $\lambda_{3,2} \geq 4$ and $\lambda_{3,3} \geq 4$. When d is a power of two, we show that $\lambda_{d,d-1} \geq d^{d/2}$ and $\lambda_{d,d} \geq d^{d/2}$. The rest of the lower bounds, we obtain by combination of the above bounds. In Section 4, we present upper bounds in $\mathbb{R}^2$ and $\mathbb{R}^3$, showing that $\lambda_{2,1} \leq 2.737$, $\lambda_{2,2} \leq 2.104$ and $\lambda_{3,3} \leq 7.807$. We conclude with open problems in Section 5.

2. Principal Component Analysis and PCA Bounding Boxes

The central idea and motivation of PCA [8] (also known as the Karhunen-Loeve transform, or the Hotelling transform) is to reduce the dimensionality of a point set by identifying *the most significant directions (principal components).* Let $X = \{x_1, x_2, \ldots, x_m\}$ be a set of vectors (points) in $\mathbb{R}^d$, and $c = (c_1, c_2, \ldots, c_d) \in \mathbb{R}^d$ be the center of gravity of X. For $1 \leq k \leq d$, we use x_{ik} to denote the k-th 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

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

[1] Preliminary results were presented in [4] and [5].

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

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

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

where C is the *covariance matrix* of X. C is a symmetric $d \times d$ matrix where the (i, j)-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. If $\chi_1 > \chi_2 > \cdots > \chi_d$ are the eigenvalues of C, then the unit eigenvector v_j for χ_j is the j-th most significant direction. All χ_js are non-negative and $\chi_j = \mathrm{var}(X, v_j)$. Since the matrix C is symmetric positive definite, its eigenvectors are orthogonal. If the eigenvalues are not distinct, the eigenvectors are not unique. In this case, an orthogonal basis of eigenvectors is chosen arbitrarily. However, we can always achieve distinct eigenvalues by a slight perturbation of the point set.

The following result summarizes the above background knowledge on PCA. For any set S of orthogonal unit vectors in $\mathbb{R}^d$, we use $\mathrm{var}(X, S)$ to denote $\sum_{v \in S} \mathrm{var}(X, v)$.

Lemma 1 *Assume that the covariance matrix C of a point set $X \in \mathbb{R}^d$ has distinct eigenvalues. For $1 \leq j \leq d$, let χ_j be the j-th largest eigenvalue of C and let v_j denote the unit eigenvector for χ_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 $\chi_1 = \max\{var(X, v) : v \in \mathbb{R}^d, \|v\| = 1\}$, and for any $2 \leq j \leq d$,*

i) $\chi_j = \max\{var(X, v) : v \in sp(B_{j-1})^\perp, \|v\| = 1\}$.

ii) $\chi_j = \min\{var(X, v) : v \in sp(B_j), \|v\| = 1\}$.

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

Since bounding boxes of a point set P (with respect to any orthogonal coordinate system) depend only on the convex hull of $CH(P)$, the construction of the covariance matrix should be based only on $CH(P)$ and not on the distribution of the points inside. Using the vertices, i.e., the 0-dimensional faces of $CH(P)$ to define the covariance matrix C we obtain a bounding box $BB_{pca(d,0)}(P)$. We denote by $\lambda_{d,0}(P)$ the approximation factor for the given point set P and by

$$\lambda_{d,0} = \sup\left\{\lambda_{d,0}(P) \mid P \subseteq \mathbb{R}^d, \mathrm{Vol}(CH(P)) > 0\right\}$$

the approximation factor in general. The example in Fig. 1 shows that $\lambda_{2,0}(P)$ can be arbitrarily large if the convex hull is a thin, slightly "bulged rectangle", with a lot of additional vertices in the middle of the two long sides. Since this construction can be lifted into higher dimensions we obtain a first general lower bound.

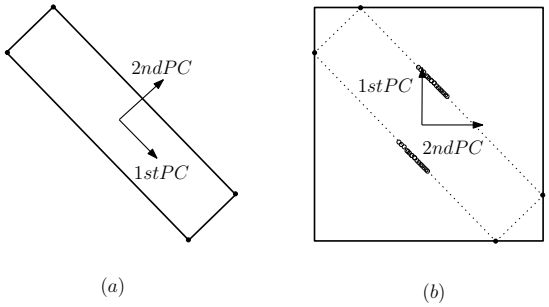


Figure 1. Four points and their PCA bounding-box (a). A dense collection of additional points significantly affect the orientation of the PCA bounding-box (b).

Proposition 1 $\lambda_{d,0} = \infty$ *for any* $d \geq 2$.

To overcome this problem, one can apply a continuous version of PCA taking into account (the dense set of) all points on the boundary of $CH(P)$, or even all points in $CH(P)$. In this approach X is a continuous set of d-dimensional vectors and the coefficients of the covariance matrix are defined by integrals instead of finite sums. If $CH(P)$ is known, the computation of the coefficients of the covariance matrix in the continuous case can also be done in linear time, thus, the overall complexity remains the same as in the discrete case. Note that for for $d = 1$ the above problem is trivial, because the PCA bounding box is always optimal, i.e., $\lambda_{1,0}$ and $\lambda_{1,1}$ are 1.

2.1. Continuous PCA

Variants of the continuous PCA applied to triangulated surfaces of 3D objects were presented by Gottschalk et. al. [6], Lahanas et. al. [10] and Vranić et. al. [16]. In what follows, we briefly review the basics of the continuous PCA in a general setting.

Let X be a continuous set of d-dimensional vectors with constant density. Then, the center of gravity of X is

$$c = \frac{\int_{x \in X} x dx}{\int_{x \in X} dx}. \tag{4}$$

Here, $\int dx$ denotes either a line integral, an area integral, or a volume integral in higher dimensions. For any unit vector $v \in \mathbb{R}^d$, the *variance of* X *in direction* v is

$$var(X, v) = \frac{\int_{x \in X} \langle x - c, v \rangle^2 dx}{\int_{x \in X} dx}. \tag{5}$$

The covariance matrix of X has the form

$$C = \frac{\int_{x \in X} (x - c)(x - c)^T dx}{\int_{x \in X} dx}, \tag{6}$$

with its (i, j)-th component

$$c_{ij} = \frac{\int_{x \in X} (x_i - c_i)(x_j - c_j) dx}{\int_{x \in X} dx}, \tag{7}$$

where x_i and x_j are the i-th and j-th component of the vector x, and c_i and c_j the i-th and j-th component of the center of gravity. It can be verified that relation (2) is also true when X is a continuous set of vectors. The procedure of finding the most significant directions can be also reformulated as an eigenvalue problem and consequently Lemma 1 holds.

For point sets P in $\mathbb{R}^2$ we are especially interested in the cases when X represents the boundary of $CH(P)$, or all points in $CH(P)$. Since the first case corresponds to the 1-dimensional faces of $CH(P)$ and the second case to the only 2-dimensional face of $CH(P)$, the generalization to a dimension $d > 2$ leads to a series of $d-1$ continuous PCA versions. For a point set $P \in \mathbb{R}^d$, $C(P, i)$ denotes the covariance matrix defined by the points on the i-dimensional faces of $CH(P)$, and $BB_{pca(d,i)}(P)$, denotes the corresponding bounding box. The approximation factors $\lambda_{d,i}(P)$ and $\lambda_{d,i}$ are defined as

$$\lambda_{d,i}(P) = \frac{Vol(BB_{pca(d,i)}(P))}{Vol(BB_{opt}(P))}, \quad \text{and}$$
$$\lambda_{d,i} = \sup\left\{\lambda_{d,i}(P) \mid P \subseteq \mathbb{R}^d, Vol(CH(P)) > 0\right\}.$$

3. Lower Bounds

The lower bounds we are going to derive are based on the following connection between the symmetry of a point set and its principal components.

Lemma 2 *Let P be a d-dimensional point set symmetric with respect to a hyperplane H and assume that the covariance matrix C has d different eigenvalues. 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 $d-1$ 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

$$C = \begin{pmatrix} c_{11} & 0 & \ldots & 0 \\ 0 & c_{22} & \ldots & c_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & c_{d2} & \ldots & c_{dd} \end{pmatrix}. \tag{8}$$

Its characteristic polynomial is

$$det(C - \chi\, I) = (c_{11} - \chi) f(\chi), \tag{9}$$

where $f(\chi)$ 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. □

We start with a generalization of Proposition 1.

Proposition 2 *$\lambda_{d,i} = \infty$ for any $d \geq 4$ and any $1 \leq i < d-1$.*

Proof. We use a lifting argument to show that for any point set $P \subseteq \mathbb{R}^k$ there is a point set $P' \subseteq \mathbb{R}^{k+1}$ such that $\lambda_{k,i}(P) \leq \lambda_{k+1,i+1}(P')$, and consequently $\lambda_{k,i} \leq \lambda_{k+1,i+1}$.

Let C be the covariance matrix of P with eigenvalues $\chi_1 > \chi_2 > \ldots > \chi_k$, and corresponding eigenvectors $v_1, v_2, \ldots v_k$. We define the point set $P'(h) = P \times \{-h, h\}, h \in \mathbb{R}^+$. Let $C'(h)$ be the covariance matrix of $P'(h)$. Obviously, the point set $P'(h)$ is symmetric with respect to the hyperplane $H = \mathbb{R}^k \times \{0\}$, and by Lemma 2, the vector $v_{k+1} = (0, \ldots, 0, 1)$ is an eigenvector of $C'(h)$. Let $\chi(h)$ be the corresponding eigenvalue of v_{k+1}. Since $\chi(h) = \mathrm{var}(P', v_{k+1})$ is a quadratic function of h, with $\lim_{h \to 0} \chi(h) = 0$, we can choose a value h_0 such that $\chi(h_0)$ is smaller than the other eigenvalues of C'. Let v be an arbitrary direction in $\mathbb{R}^k$. Then, by definition of P', the variance of P' in the direction $(v, 0)$ remains the same as the variance of P in the direction v. Thus, we can conclude that the eigenvalues of C' are $\chi_1 > \chi_2 > \ldots > \chi_k > \chi(h_0)$, with corresponding eigenvectors $(v_1, 0), (v_2, 0), \ldots (v_k, 0), v_{k+1}$, and consequently $Vol(BB_{pca(k+1,i+1)}(P')) = 2\, h_0\, Vol(BB_{pca(k,i)}(P))$.

On the other hand, the bounding box $BB_{h_0} = BB_{opt}(P) \times [-h_0, h_0]$ is also a bounding box of P'. Therefore, we obtain

$$\lambda_{k+1,i+1} \geq \lambda_{k+1,i+1}(P') = \frac{Vol(BB_{pca(k+1,i+1)}(P'))}{Vol(BB_{opt}(P'))} \geq \frac{Vol(BB_{pca(k+1,i+1)}(P'))}{Vol(BB_{h_0})}$$
$$\geq \frac{2h_0 Vol(BB_{pca(k,i)}(P))}{2h_0 Vol(BB_{opt}(P))} \geq \lambda_{k,i}.$$

Now, we can establish $\lambda_{d,i} \geq \lambda_{d-1,i-1} \geq \ldots \geq \lambda_{d-i,0} = \infty$. □

This way, there remain only two interesting cases for a given d: the factor $\lambda_{d,d-1}$ corresponding to the boundary of the convex hull, and the factor $\lambda_{d,d}$ corresponding to the full convex hull.

3.1. Lower bounds in $\mathbb{R}^2$

The result obtained in this subsection can be seen as a special case of the result obtained in Subsection 3.3. To gain a better understanding of the problem and the obtained results, we consider it separately.

Theorem 1 *$\lambda_{2,1} \geq 2$ and $\lambda_{2,2} \geq 2$.*

Proof. Both lower bounds can be derived from a rhombus. Let the side length of the rhombus be 1. To make sure that the covariance matrix has two distinct eigenvalues, we assume that the rhombus has an angle $\alpha > 90°$. Since the rhombus is symmetric, its PCs coincide with its diagonals. In Fig. 2 (b) its optimal-area bounding boxes, for 2 different angles, $\alpha > 90°$ and $\beta = 90°$, are shown, and in Fig. 2 (a) its corresponding PCA bounding boxes. As the rhombus' angles in limit approach 90°, the rhombus approaches a square with side length 1, i.e., the vertices of the rhombus in the limit are $(\frac{1}{\sqrt{2}}, 0), (-\frac{1}{\sqrt{2}}, 0), (0, \frac{1}{\sqrt{2}})$ and $(0, -\frac{1}{\sqrt{2}})$ (see Fig. 2 (a)), and the dimensions of its PCA bounding box are $\sqrt{2} \times \sqrt{2}$. According to Lemma 2, the PCs of the rhombus are unique as long its angles are not 90°. This leads to the conclusion that the ratio between the area of the PCA bounding box in Fig. 2 (a), and the area of the optimal-area bounding box in Fig. 2 (b), in limit goes to 2. □

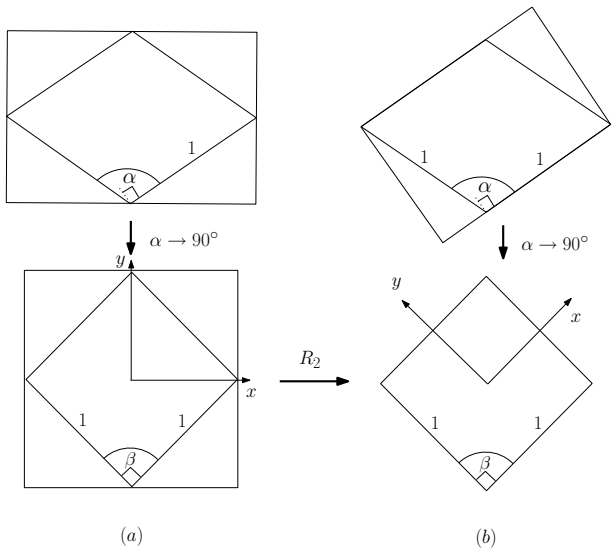


Figure 2. An example which gives the lower bound of the area of the PCA bounding box of an arbitrary convex polygon in $\mathbb{R}^2$.

Alternatively, to show that the given squared rhombus fits into a unit cube, one can apply the following rotation matrix

$$R_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \tag{10}$$

It can be verified easily that all coordinates of the vertices of the rhombus transformed by R_2 are in the interval $[-0.5, 0.5]$. We use similar arguments when we prove the lower bounds in higher dimensions.

3.2. Lower bounds in $\mathbb{R}^3$

Theorem 2 *$\lambda_{3,2} \geq 4$ and $\lambda_{3,3} \geq 4$.*

Proof. Both lower bounds are obtained from a dipyramid, having a rhombus with side length $\sqrt{2}$ as its base. The other sides of the dipyramid have length $\frac{\sqrt{3}}{2}$. Similarly as in $\mathbb{R}^2$, we consider the case when its base, the rhombus, in limit approaches the square, i.e., the vertices of the square dipyramid are $(1,0,0), (-1,0,0), (0,1,0), (0,-1,0), (0,0,\frac{\sqrt{2}}{2})$ and $(0,0,-\frac{\sqrt{2}}{2})$ (see Fig. 3 (a)). The dimensions of its PCA bounding box are $2 \times 2 \times \sqrt{2}$. Now, we rotate the coordinate system (or the square dipyramid) with the rotation determined by the following orthogonal matrix

$$R_3 = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} \end{bmatrix}. \tag{11}$$

It can be verified easily that the square dipyramid, after rotation with R_3 fits into the box $[-0.5, 0.5]^3$ (see Fig. 3 (b)). Thus, the ratio of the volume of the bounding box, Fig. 3 (a), and the volume of its PCA bounding box, Fig. 3 (b), in limit goes to 4. □

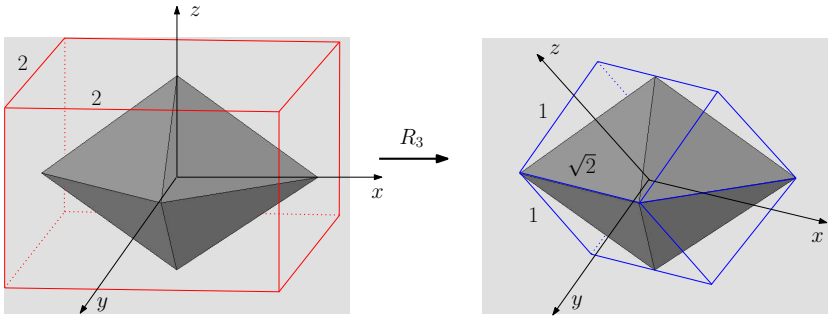


Figure 3. An example which gives the lower bound of the volume of the PCA bounding box of an arbitrary convex polygon in $\mathbb{R}^3$.

3.3. Lower bounds in $\mathbb{R}^d$

Theorem 3 *If d is a power of two, then $\lambda_{d,d-1} \geq d^{d/2}$ and $\lambda_{d,d} \geq d^{d/2}$.*

Proof. For any $d = 2^k$, let a_i be a d-dimensional vector, with $a_{ii} = \frac{\sqrt{d}}{2}$ and $a_{ij} = 0$ for $i \neq j$, and let $b_i = -a_i$. We construct a d-dimensional convex polytope P_d with vertices $V = \{a_i, b_i | 1 \leq i \leq d\}$. It is easy to check that the hyperplane normal to a_i is a hyperplane of reflective symmetry, and as consequence of Lemma 2, a_i is an eigenvector of the covariance matrix of P_d. To ensure that all eigenvalues are different (which implies that the PCA bounding box is unique), we add $\epsilon_i > 0$ to the i-th coordinate of a_i, and $-\epsilon_i$ to the i-th coordinate of b_i, for $1 \leq i \leq d$, where $\epsilon_1 < \epsilon_2 < \ldots < \epsilon_d$. When all ϵ_i, $1 \leq i \leq d$, arbitrary approach 0, the PCA bounding box of the convex polytope P_d converges to a hypercube with side lengths $\sqrt{d}$, i.e., the volume of the PCA bounding box of P_d converges to $d^{d/2}$. Now, we rotate P_d, such that it fits into the cube $[-\frac{1}{2}, \frac{1}{2}]^d$. For $d = 2^k$, we can use a rotation matrix derived from a *Hadamard matrix*[2], recursively defined by

$$R_d = \frac{1}{\sqrt{2}} \left[\begin{array}{c|c} R_{\frac{d}{2}} & R_{\frac{d}{2}} \\ \hline R_{\frac{d}{2}} & -R_{\frac{d}{2}} \end{array} \right], \tag{12}$$

where we start with the matrix R_2 defined above (10) for $d = 2$. A straightforward calculation verifies that P_d rotated with R_d fits into the cube $[-0.5, 0.5]^d$. □

Remark: Theorem 3 holds for all dimensions d for which a $d \times d$ Hadamard matrix exists. As it was shown in the proof of the theorem, this is always true when d is a power of two. Moreover, Hadamard conjectured that a $d \times d$ Hadamard matrix exists when d is a multiple of four. This conjecture is known to be true for $d \leq 664$ [9].

We can combine lower bounds from lower dimensions to get lower bounds in higher dimensions by taking Cartesian products. If λ_{d_1} is a lower bound on the ratio between

[2] A Hadamard matrix is a ± 1 matrix with orthogonal columns.

the PCA bounding box and the optimal bounding box of a convex polytope in $\mathbb{R}^{d_1}$, and λ_{d_2} is a lower bound in $\mathbb{R}^{d_2}$, then $\lambda_{d_1} \cdot \lambda_{d_2}$ is a lower bound in $\mathbb{R}^{d_1+d_2}$. This observation together with the results from this section enables us to obtain lower bounds in any dimension. For example, for the first 10 dimensions, the lower bounds we obtain are given in Table 1.

Table 1
Lower bounds for the approximation factor of PCA bounding boxes for the first 10 dimensions.

dimension	$\mathbb{R}$	$\mathbb{R}^2$	$\mathbb{R}^3$	$\mathbb{R}^4$	$\mathbb{R}^5$	$\mathbb{R}^6$	$\mathbb{R}^7$	$\mathbb{R}^8$	$\mathbb{R}^9$	$\mathbb{R}^{10}$
lower bound	1	2	4	16	16	32	64	4096	4096	8192

4. Upper Bounds

4.1. An upper Bound on $\lambda_{2,1}$

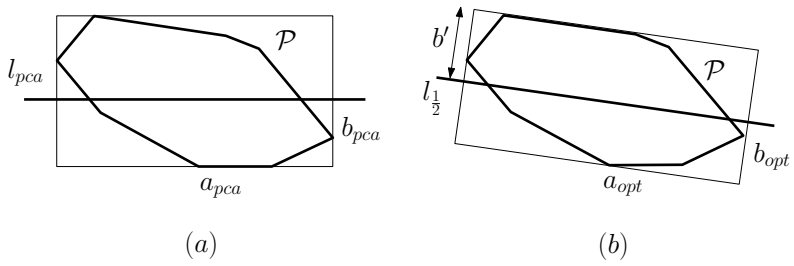
Given a point set $P \subseteq \mathbb{R}^2$ and an arbitrary bounding box $BB(P)$ we will denote the two side lengths by a and b, where $a \geq b$. We are interested in the side lengths $a_{opt}(P) \geq b_{opt}(P)$ and $a_{pca}(P) \geq b_{pca}(P)$ of $BB_{opt}(P)$ and $BB_{pca(2,1)}(P)$, see Fig. 4. The parameters $\alpha = \alpha(P) = a_{pca}(P)/a_{opt}(P)$ and $\beta = \beta(P) = b_{pca}(P)/b_{opt}(P)$ denote the ratios between the corresponding side lengths. Hence, we have $\lambda_{2,1}(P) = \alpha(P) \cdot \beta(P)$. If the relation to P is clear, we will omit the reference to P in the notations introduced above.

Since the side lengths of any bounding box are bounded by the diameter of P, we can observe that in general $b_{pca}(P) \leq a_{pca}(P) \leq diam(P) \leq \sqrt{2}a_{opt}(P)$, and in the special case when the optimal bounding box is a square $\lambda_{2,1}(P) \leq 2$. This observation can be generalized, introducing an additional parameter $\eta(P) = a_{opt}(P)/b_{opt}(P)$.

Lemma 3 *$\lambda_{2,1}(P) \leq \eta + \frac{1}{\eta}$ and $\lambda_{2,2}(P) \leq \eta + \frac{1}{\eta}$ for any point set P with fixed aspect ratio $\eta(P) = \eta$.*

Proof. We have for both a_{pca} and b_{pca} the upper bound $diam(P) \leq \sqrt{a_{opt}^2 + b_{opt}^2} = a_{opt}\sqrt{1 + \frac{1}{\eta^2}}$. Replacing a_{opt} by $\eta \cdot b_{opt}$ in the bound on b_{pca} we obtain $\alpha\beta \leq \eta \left(\sqrt{1 + \frac{1}{\eta^2}}\right)^2 = \eta + \frac{1}{\eta}$. □

Unfortunately, this parametrized upper bound tends to infinity for $\eta \to \infty$. Therefore, we are going to derive another upper bound that is better for large values of η. In this process we will make essential use of the properties of $BB_{pca(2,1)}(P)$. In order to distinguish clearly between a convex set and its boundary, we will use calligraphic letters for the boundaries, specifically $\mathcal{P}$ for the boundary of $CH(P)$ and $\mathcal{BB}_{opt}$ for the boundary of the rectangle $BB_{opt}(P)$. Furthermore, we denote by $d^2(\mathcal{P}, l)$ the integral of the squared distances of the points on $\mathcal{P}$ to a line l, i.e., $d^2(\mathcal{P}, l) = \int_{x \in \mathcal{P}} d^2(x, l) ds$. Let l_{pca} be the line going through the center of gravity and parallel to the longer side of $BB_{pca(2,1)}(P)$ and $l_{\frac{1}{2}}$ be the bisector



(a) (b)

Figure 4. A convex polygon $\mathcal{P}$, its PCA bounding box and the line l_{pca}, which coincides with the first principal component of $\mathcal{P}$ (a). The optimal bounding box and the line $l_{\frac{1}{2}}$, going through the middle of its smaller side, parallel with its longer side (b).

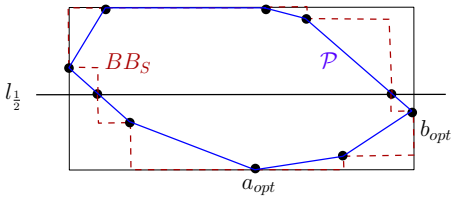


Figure 5. The convex polygon $\mathcal{P}$, its optimal bounding box, and the staircase polygon BB_S (depicted dashed).

of $BB_{opt(P)}$ parallel to the longer side. By Lemma 1, part ii) l_{pca} is the best fitting line of P and therefore,

$$d^2(\mathcal{P}, l_{pca}) \leq d^2(\mathcal{P}, l_{\frac{1}{2}}). \tag{13}$$

Lemma 4 $d^2(\mathcal{P}, l_{\frac{1}{2}}) \leq \frac{{b_{opt}}^2 a_{opt}}{2} + \frac{{b_{opt}}^3}{6}$.

Proof. If a segment of $\mathcal{P}$ intersects the line $l_{\frac{1}{2}}$, we split this segment into two segments, with the intersection point as a split point. Then, to each segment f of $\mathcal{P}$ flush with the side of the PCA bounding box, we assign a segment identical to f. To each remaining segment s of $\mathcal{P}$, with endpoints (x_1, y_1) and (x_2, y_2), where $|y_1| \leq |y_2|$, we assign two segments: a segment s_1, with endpoints (x_1, y_1) and (x_1, y_2), and a segment s_2, with endpoints (x_1, y_2) and (x_2, y_2). All these segments form the boundary $\mathcal{BB}_S$ of a staircase polygon (see Fig. 5 for illustration). Two straightforward consequences are that $d^2(\mathcal{BB}_S, l_{\frac{1}{2}}) \leq d^2(\mathcal{BB}_{opt}, l_{\frac{1}{2}})$, and $d^2(s, l_{\frac{1}{2}}) \leq d^2(s_1, l_{\frac{1}{2}}) + d^2(s_2, l_{\frac{1}{2}})$, for each segment s of $\mathcal{P}$. Therefore, $d^2(\mathcal{P}, l_{\frac{1}{2}})$ is at most $d^2(\mathcal{BB}_S, l_{\frac{1}{2}})$, which is bounded from above by $d^2(\mathcal{BB}_{opt}, l_{\frac{1}{2}}) = 4\int_0^{\frac{b_{opt}}{2}} x^2\,dx + 2\int_0^{a_{opt}} (\frac{b_{opt}}{2})^2\,dx = \frac{{b_{opt}}^2 a_{opt}}{2} + \frac{{b_{opt}}^3}{6}$. □

Now we look at $\mathcal{P}$ and its PCA bounding box (Fig. 6). The line l_{pca} divides $\mathcal{P}$ into an upper and a lower part, $\mathcal{P}_{upp}$ and $\mathcal{P}_{low}$. l_{upp} denotes the orthogonal projection of $\mathcal{P}_{upp}$

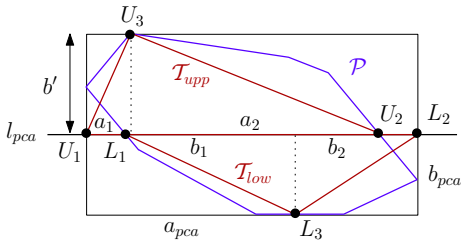


Figure 6. The convex polygon $\mathcal{P}$, its PCA bounding box, and a construction for a lower bound on $d^2(\mathcal{P}, l_{pca})$

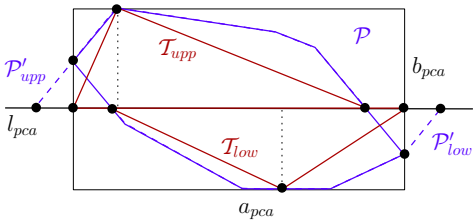


Figure 7. Two polylines $\mathcal{P}'_{upp}$ and $\mathcal{P}'_{low}$ (depicted dashed) formed from $\mathcal{P}$.

onto l_{pca}, with U_1 and U_2 as its extreme points, and l_{low} denotes the orthogonal projection of $\mathcal{P}_{low}$ onto l_{pca}, with L_1 and L_2 as its extreme points. $\mathcal{T}_{upp} = \triangle(U_1U_2U_3)$ is a triangle inscribed in $\mathcal{P}_{upp}$, where point U_3 lies on the intersection of $\mathcal{P}_{upp}$ with the upper side of the PCA bounding box. Analogously, $\mathcal{T}_{low} = \triangle(L_1L_2L_3)$ is a triangle inscribed in $\mathcal{P}_{low}$.

Lemma 5 $d^2(\mathcal{P}, l_{pca}) \geq d^2(\mathcal{T}_{upp}, l_{pca}) + d^2(\mathcal{T}_{low}, l_{pca})$.

Proof. Let Q denote a chain of segments of $\mathcal{P}$, which does not touch the longer side of the PCA bounding box, and whose one endpoint lies on the smaller side of the PCA bounding box, and the other endpoint on the line l_{pca}. We reflect Q at the line supporting the side of the PCA bounding box touched by Q. All such reflected chains of segments, together with the rest of $\mathcal{P}$, form two polylines: $\mathcal{P}'_{upp}$ and $\mathcal{P}'_{low}$ (see Fig. 7 for illustration). As a consequence, to each of the sides of the triangles $\mathcal{T}_{low}$ and $\mathcal{T}_{upp}$, $\overline{L_1L_3}$, $\overline{L_2L_3}$, $\overline{U_1U_3}$, $\overline{U_2U_3}$, we have a corresponding chain of segments R as shown in the two cases in Fig. 8. In both cases $d^2(t, l_{pca}) \leq d^2(R, l_{pca})$. Namely, we can parametrize both curves, R and t, starting at the common endpoint A that is furthest from l_{pca}. By comparing two points with the same parameter (distance from A along the curve) we see that the point on t always has a smaller distance to l_{pca} than the corresponding point on R. In addition t is shorter, and some parts of R have no match on t.

Consequently, $d^2(\mathcal{P}', l_{pca}) \geq d^2(\mathcal{T}_{upp} \bigcup \mathcal{T}_{low}, l_{pca}) = d^2(\mathcal{T}_{upp}, l_{pca}) + d^2(\mathcal{T}_{low}, l_{pca})$, and since $d^2(\mathcal{P}', l_{pca}) = d^2(\mathcal{P}, l_{pca}) = d^2(\mathcal{P}_{upp} \bigcup \mathcal{P}_{low}, l_{pca})$, the proof is completed. □

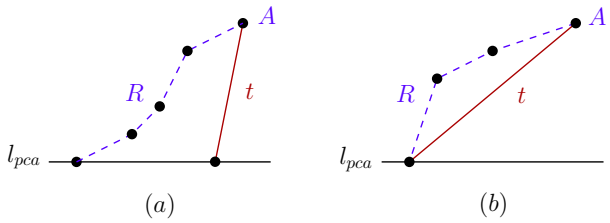


Figure 8. Two types of chains of segments (depicted dashed and denoted by R), and their corresponding triangles' edges (depicted solid and denoted by t). The base-point of t corresponds to the most left point of $\mathcal{T}_{upp}$ from Fig. 6 and Fig. 7.

Since $\mathcal{P}$ is convex, the following relations hold:

$$|l_{upp}| \geq \frac{b'}{b_{pca}} a_{pca}, \text{ and } |l_{low}| \geq \frac{b_{pca} - b'}{b_{pca}} a_{pca}. \tag{14}$$

The value

$$\begin{aligned} d^2(\mathcal{T}_{upp}, l_{pca}) &= \int_0^{\sqrt{a_1^2+b'^2}} (\tfrac{\alpha}{\sqrt{a_1^2+b'^2}} b')^2 \, d\alpha \; + \; \int_0^{\sqrt{a_2^2+b'^2}} (\tfrac{\alpha}{\sqrt{a_2^2+b'^2}} b')^2 \, d\alpha \\ &= \tfrac{b'^2}{3} (\sqrt{a_1^2+b'^2} + \sqrt{a_2^2+b'^2}) \end{aligned}$$

is minimal when $a_1 = a_2 = \frac{|l_{upp}|}{2}$. With (14) 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 Lemma 5, gives:

$$d^2(\mathcal{P}, l_{pca}) \geq \frac{b_{pca}^2}{12} \sqrt{a_{pca}^2 + 4b_{pca}^2}. \tag{15}$$

Combining (13), (15) and Lemma 4 we have:

$$\frac{1}{2} a_{opt} b_{opt}^2 + \frac{1}{6} b_{opt}^3 \geq \frac{b_{pca}^2}{12} \sqrt{a_{pca}^2 + 4b_{pca}^2} \geq \frac{b_{pca}^2}{12} a_{pca}. \tag{16}$$

Replacing a_{opt} with ηb_{opt} on the left side, b_{pca}^2 with $\beta^2 b_{opt}^2$ and a_{pca} with $\alpha a_{opt} = \alpha\eta b_{opt}$ on the right side of (16), we obtain:

$$\left(\frac{\eta}{2} + \frac{1}{6}\right) b_{opt}^3 \geq \frac{\beta^2 \, \alpha \, \eta}{12} b_{opt}^3$$

which implies

$$\beta \leq \sqrt{\frac{6\eta+2}{\alpha\,\eta}}.$$

This gives the second upper bound on $\lambda_{2,1}(P)$ for point sets with parameter η:

$$\alpha\,\beta \leq \sqrt{\frac{(6\eta+2)\alpha}{\eta}} \leq \sqrt{\frac{6\eta+2}{\eta}\sqrt{1+\frac{1}{\eta^2}}}. \tag{17}$$

Lemma 6 $\lambda_{2,1}(P) \leq \sqrt{\frac{6\eta+2}{\eta}\sqrt{1+\frac{1}{\eta^2}}}$ *for any point set P with fixed aspect ratio $\eta(P)=\eta$.*

This implies the final result of this subsection.

Theorem 4 *The PCA bounding box of a point set P in $\mathbb{R}^2$ computed over the boundary of $CH(P)$ has a guaranteed approximation factor $\lambda_{2,1} \leq 2.737$.*

Proof. The theorem follows from the combination of the two parametrized bounds from Lemma 3 and Lemma 6 proved above:

$$\lambda_{2,1} \leq \sup_{\eta\geq 1}\left\{\min\left(\eta+\frac{1}{\eta},\ \sqrt{\frac{6\eta+2}{\eta}\sqrt{1+\frac{1}{\eta^2}}}\right)\right\}.$$

It is easy to check that the supremum $s \approx 2.736$ is obtained for $\eta \approx 2.302$. □

Although this result concerns a continuous PCA version, the proof is mainly based on arguments from discrete geometry. In contrast to that, the upper bound proofs for $\lambda_{2,2}$ and $\lambda_{3,3}$, presented in the next two subsections, essentially make use of integral calculus.

4.2. An upper bound on $\lambda_{2,2}$

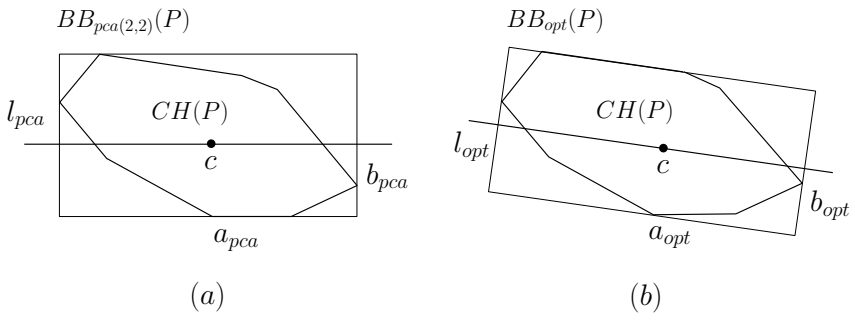


Figure 9. A convex hull of the point set P, its PCA bounding box (a) and its optimal bounding box (b).

First, we note that due to Lemma 3, we already have a parametrized upper bound on $\lambda_{2,2}$. Since this bound tends to infinity for $\eta \to \infty$, we are going to derive another

upper bound on $\lambda_{2,2}$ that is better for large values of η. We derive such a bound by finding a constant that bounds β from above. In this process we will make essential use of the properties of $BB_{pca(2,2)}(P)$. We denote by $d^2(CH(P), l)$ the integral of the squared distances of the points on $CH(P)$ to a line l, i.e.,

$$d^2(CH(P), l) = \int_{s \in CH(P)} d^2(s, l) ds.$$

Let l_{pca} be the line going through the center of gravity, parallel to the longer side of $BB_{pca(2,2)}(P)$, and l_{opt} be the line going through the center of gravity, parallel to the longer side of $BB_{opt(P)}$ (see Fig. 9). By Lemma 1, part *ii*) l_{pca} is the best fitting line of P and therefore,

$$d^2(CH(P), l_{pca}) \leq d^2(CH(P), l_{opt}). \tag{18}$$

We obtain an estimate of β by determining a lower bound on $d^2(CH(P), l_{pca})$ that depends on b_{pca}, and an upper bound on $d^2(CH(P), l_{opt})$ that depends on b_{opt}. Having an arbitrary bounding box of $CH(P)$ (with side lengths a and b, $a \geq b$) the area of $CH(P)$ can be expressed as

$$A = A(CH(P)) = \int_0^b \int_0^a \chi_{CH(P)}(x, y) dx dy = \int_0^b g(y) dy,$$

where $\chi_{CH(P)}(x, y)$ is the *characteristic function* of $CH(P)$ defined as

$$\chi_{CH(P)}(x, y) = \begin{cases} 1 & (x, y) \in CH(P) \\ 0 & (x, y) \notin CH(P), \end{cases}$$

and $g(y) = \int_0^a \chi_{CH(P)}(x, y) dx$ is the length of the intersection of $CH(P)$ with a horizontal line at height y. In the following we call $g(y)$ the *density function* of $CH(P)$ for computing the area with the integral $\int_0^b g(y) dy$. Since $CH(P)$ is a convex set, $g(y)$ is continuous and convex in the interval $[0, b]$ (see Fig. 10 (a) for an illustration). Let b_1 denote the y-coordinate of the center of gravity of $CH(P)$. The line l_{b_1} ($y = b_1$) divides the area of $CH(P)$ into A_1 and A_2.

Theorem 6, which is derived from the generalized first mean value theorem of integral calculus (Theorem 5), is our central technical tool in derivation of the lower and the upper bound on $d^2(CH(P), l_{b_1})$.

Theorem 5 (Generalized first mean value theorem of integral calculus)
If $h(x)$ and $g(x)$ are continuous functions on the interval $[a, b]$, and if $g(x)$ does not change its sign in the interval, then there is a $\xi \in (a, b)$ such that

$$\int_a^b h(x) g(x) dx = h(\xi) \int_a^b g(x) dx.$$

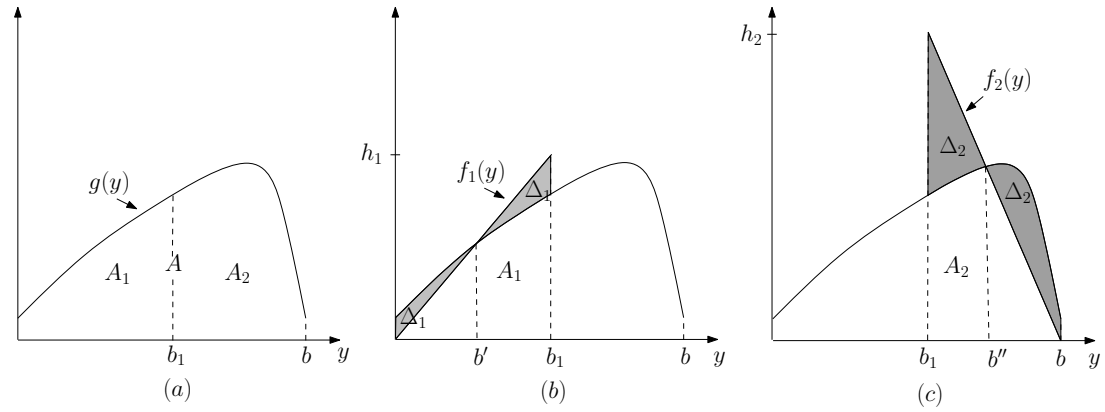


Figure 10. Construction of the lower bound on $d^2(CH(P), l_{b_1})$.

Theorem 6 *Let $f(x)$ and $g(x)$ be positive continuous functions on the interval $[a, b]$ with $\int_a^b f(x)dx = \int_a^b g(x)dx$, and assume that there is some $c \in [a, b]$ such that $f(x) \leq g(x)$, for all $x \leq c$ and $f(x) \geq g(x)$, for all $x \geq c$. Then*

$$\int_a^b (x-b)^2 f(x)dx \leq \int_a^b (x-b)^2 g(x)dx \quad \text{and}$$

$$\int_a^b (x-a)^2 f(x)dx \geq \int_a^b (x-a)^2 g(x)dx.$$

Proof. We start from the assumptions $\int_a^b f(x)dx = \int_a^b g(x)dx$ and $f(x) \leq g(x)$ for all $x \leq c$ and $f(x) \geq g(x)$ for all $x \geq c$. Thus,

$$\int_a^c (g(x) - f(x))dx = \int_c^b (f(x) - g(x))dx = \Delta \tag{19}$$

and the integrands on both sides are nonnegative. Applying Theorem 5 to the following integrals we obtain

$$\int_a^c (x-b)^2 (g(x) - f(x))dx \;=\; (\xi_1 - b)^2 \int_a^c (g(x) - f(x))dx \;=\; (\xi_1 - b)^2 \Delta,$$

and

$$\int_c^b (x-b)^2 (f(x) - g(x))dx \;=\; (\xi_2 - b)^2 \int_c^b (f(x) - g(x))dx \;=\; (\xi_2 - b)^2 \Delta,$$

for some $\xi_1 \in [a, c]$ and $\xi_2 \in [c, b]$. Therefore,

$$\int_a^c (x-b)^2 (g(x) - f(x))dx = (\xi_1 - b)^2 \Delta \geq (\xi_2 - b)^2 \Delta = \int_c^b (x-b)^2 (f(x) - g(x))dx.$$

It follows that

$$\int_a^b (x-b)^2(g(x)-f(x))dx = \int_a^c (x-b)^2(g(x)-f(x))dx - \int_c^b (x-b)^2(f(x)-g(x))dx \geq 0,$$

which proves the first claim

$$\int_a^b (x-b)^2 f(x)dx \leq \int_a^b (x-b)^2 g(x)dx.$$

The proof of the second claim follows by symmetry. □

The following theorem was discovered independently by Grünbaum [7] and Hammer (unpublished manuscript), and later rediscovered by Mityagin [11]. We use it to prove a lower and an upper bound of the variance $d^2(CH(P), l_{b_1})$.

Theorem 7 (Grünbaum-Hammer-Mityagin) *Let K be a compact convex set in $\mathbb{R}^d$ with non-empty interior and centroid μ. Assume that the d-dimensional volume of K is one, that is, $Vol_d(K) = 1$. Let H be any (d-1)-dimensional hyperplane passing through μ with corresponding half-spaces H^+ and H^-. Then,*

$$\min\{\, Vol_d(K \cap H^+),\, Vol_d(K \cap H^-)\} \geq \left(\frac{d}{d+1}\right)^d$$

Moreover, the bound $(\frac{d}{d+1})^d$ is best possible.

Lemma 7 *The variance $d^2(CH(P), l_{b_1})$ is bounded from below by $\frac{10}{243}Ab^2$.*

Proof. We split the integral $\int_0^b (y-b_1)^2 g(y)dy$ at b_1 (recall that b_1 is the y-coordinate of the center of gravity of $CH(P)$), and prove lower bounds on both parts in the following way: For the left part consider the linear function $f_1(y) = \frac{h_1}{b_1}y$ such that $\int_0^{b_1} f_1(y)dy = \int_0^{b_1} g(y)dy = A_1$ (see Fig. 10 (b) for an illustration). From $\int_0^{b_1} f_1(y)dy = A_1$, it follows that $f_1(y) = \frac{2A_1 y}{b_1^2}$. Since $g(y)$ is convex, $g(y)$ and $f_1(y)$ intersect only once, at a point $b' \in (0, b_1)$. By Theorem 6, we have

$$\int_0^{b_1} (y-b_1)^2 g(y)dy \geq \int_0^{b_1} (y-b_1)^2 f_1(y)dy = \int_0^{b_1} (y-b)^2 \frac{2A_1}{{b_1}^2}dy = \frac{A_1 b_1^2}{6}. \tag{20}$$

Analogously, for the right part consider the linear function $f_2(y) = \frac{h_2}{b_1-b}(y-b) = \frac{h_2}{-b_2}(y-b)$ such that $\int_{b_1}^b f_2(y)dy = \int_{b_1}^b g(y)dy = A_2$ (see Fig. 10 (c) for an illustration). From $\int_{b_1}^b f_2(y)dy = A_2$, it follows that $f_2(y) = \frac{2A_2}{b_2^2}(y-b)$. Since $g(y)$ is convex, $g(y)$ and $f_2(y)$ intersect only once, at a point $b'' \in (b_1, b)$. By Theorem 6, we have that

$$\begin{aligned} \int_{b_1}^b (y-b_1)^2 g(y)dy &\geq \int_{b_1}^b (y-b_1)^2 f_2(y)dy = \int_{b_1}^b (y-b_1)^2 \tfrac{2A_2}{(b-b_1)^2}(y-b_1)dy \\ &= \tfrac{A_2 b_2^2}{6}. \end{aligned} \tag{21}$$

From (20) and (21) we obtain that

$$d^2(CH(P), l_{b_1}) = \int_0^{b_1}(y-b_1)^2 g(y)dy + \int_{b_1}^{b}(y-b_1)^2 g(y)dy \geq \frac{A_1 b_1^2}{6} + \frac{A_2 b_2^2}{6}.$$

From the Grünbaum-Hammer-Mityagin theorem, we know that $A_1, A_2 \in [\frac{4}{9}A, \frac{5}{9}A]$. Also, we know that $b_1, b_2 \in [\frac{1}{3}b, \frac{2}{3}b]$. It is not hard to show that, under these constrains, the expression $\frac{A_1 b_1^2}{6} + \frac{A_2 b_1^2}{6}$ achieves its minimum of $\frac{10}{243}Ab^2$ for $A_1 = \frac{4}{9}A, b_1 = \frac{5}{9}b$ or $A_1 = \frac{5}{9}A, b_1 = \frac{4}{9}b$. □

Lemma 8 *The variance $d^2(CH(P), l_{b_1})$ is bounded from above by $\frac{29}{243}Ab^2$.*

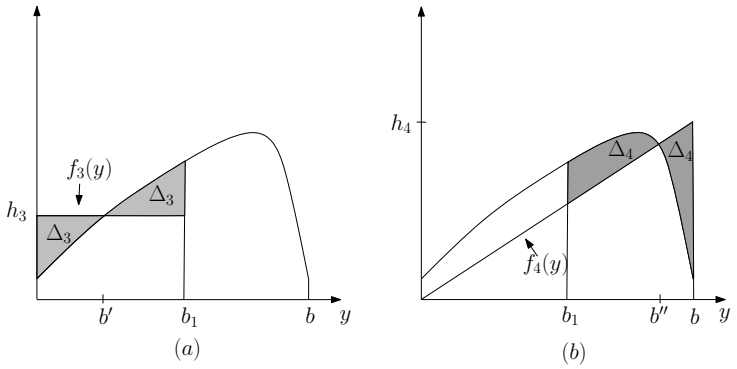


Figure 11. Construction of the upper bound on $d^2(CH(P), l_{b_1})$.

The proof of Lemma 8 is similar to the proof of Lemma 7. Here, the functions we use to derive the upper bound on $d^2(CH(P), l_{b_1})$ are given in Fig 11 (functions $f_3(y)$ and $f_4(y)$).

Now, we are ready to derive an alternative parametrized upper bound on $\lambda_{2,2}(P)$ which is better than the bound from Lemma 3 for big values of η.

Lemma 9 $\lambda_{2,2}(P) \leq \sqrt{2.9\left(1+\frac{1}{\eta^2}\right)}$ *for any point set P with aspect ratio $\eta(P) = \eta$.*

Proof. Applying Lemma 7 and Lemma 8 in (18) we obtain

$$\frac{10}{243}Ab_{pca}^2 \leq d^2(\mathcal{P}, l_{pca}) \leq d^2(\mathcal{P}, l_{opt}) \leq \frac{29}{243}Ab_{opt}^2. \tag{22}$$

From (22) it follows that $\beta = \frac{b_{pca}}{b_{opt}} \leq \sqrt{2.9}$. We have for a_{pca} the upper bound $diam(P) \leq \sqrt{a_{opt}^2 + b_{opt}^2} = a_{opt}\sqrt{1+\frac{1}{\eta^2}}$. From this, it follows that $\alpha \leq \sqrt{1+\frac{1}{\eta^2}}$. Putting this together, we obtain $\alpha\beta \leq \sqrt{2.9\left(1+\frac{1}{\eta^2}\right)}$. □

Theorem 8 *The PCA bounding box of a point set P in $\mathbb{R}^2$ computed over $CH(P)$ has a guaranteed approximation factor $\lambda_{2,2} \leq 2.104$.*

Proof. The theorem follows from the combination of the two parametrized bounds from Lemma 3 and Lemma 9:

$$\lambda_{2,2} \leq \sup_{\eta \geq 1} \left\{ \min \left(\eta + \frac{1}{\eta}, \sqrt{2.9 \left(1 + \frac{1}{\eta^2} \right)} \right) \right\}.$$

It is easy to check that the supremum $s \approx 2.1038$ is obtained for $\eta \approx 1.3784$. □

4.3. An upper bound on $\lambda_{3,3}$

Some of the techniques used here are similar to those used in Subsection 4.2 where we derive an upper bound on $\lambda_{2,2}$. One essential difference is that for the upper bound on $\lambda_{3,3}$, we additionally need a bound for the ratio of the middle sides of $BB_{pca(3,3)}(P)$ and $BB_{opt}(P)$, which we derive from the relation in Lemma 13.

Given a point set $P \subseteq \mathbb{R}^3$ and an arbitrary bounding box $BB(P)$, we will denote the three side lengths of $BB(P)$ by a,b and c, where $a \geq b \geq c$. We are interested in the side lengths $a_{opt} \geq b_{opt} \geq c_{opt}$ and $a_{pca} \geq b_{pca} \geq c_{pca}$ of $BB_{opt}(P)$ and $BB_{pca(3,3)}(P)$. The parameters $\alpha = \alpha(P) = a_{pca}/a_{opt}$, $\beta = \beta(P) = b_{pca}/b_{opt}$ and $\gamma = \gamma(P) = c_{pca}/c_{opt}$ denote the ratios between the corresponding side lengths. Hence, we have $\lambda_{3,3}(P) = \alpha \cdot \beta \cdot \gamma$.

Since the side lengths of any bounding box are bounded by the diameter of P, we can observe that in general $c_{pca} \leq b_{pca} \leq a_{pca} \leq diam(P) \leq \sqrt{3} a_{opt}$, and in the special case when the optimal bounding box is a cube $\lambda_{3,3}(P) \leq 3\sqrt{3}$. This observation can be generalized, introducing two additional parameters $\eta(P) = a_{opt}/b_{opt}$ and $\theta(P) = a_{opt}/c_{opt}$.

Lemma 10 $\lambda_{3,3}(P) \leq \eta\,\theta \left(1 + \frac{1}{\eta^2} + \frac{1}{\theta^2}\right)^{\frac{3}{2}}$ *for any point set P with aspect ratios $\eta(P) = \eta$ and $\theta(P) = \theta$.*

Proof. We have for a_{pca}, b_{pca} and c_{pca} the upper bound $diam(P) \leq \sqrt{a_{opt}^2 + b_{opt}^2 + c_{opt}^2}$ $= a_{opt}\sqrt{1 + \frac{1}{\eta^2} + \frac{1}{\theta^2}}$. Thus, $\alpha\,\beta\,\gamma \leq \frac{a_{pca}\,b_{pca}\,c_{pca}}{a_{opt}\,b_{opt}\,c_{opt}} \leq \frac{a_{opt}^3 \left(1+\frac{1}{\eta^2}\right)^{\frac{3}{2}}}{a_{opt} b_{opt} c_{opt}}$. Replacing a_{opt} in the nominator once by $\eta\, b_{opt}$ and once by $\theta\, c_{opt}$ we obtain $\lambda_{3,3}(P) \leq \eta\,\theta \left(1 + \frac{1}{\eta^2} + \frac{1}{\theta^2}\right)^{\frac{3}{2}}$. □

Unfortunately, this parametrized upper bound tends to infinity for $\eta \to \infty$ or $\theta \to \infty$. Therefore, we are going to derive another upper bound that is better for large values of η and θ. We derive such a bound by finding constants that bound β and γ from above. In this process we will make essential use of the properties of $BB_{pca(3,3)}(P)$. We denote by $d^2(CH(P), H)$ the integral of the squared distances of the points on $CH(P)$ to a plane H, i.e., $d^2(CH(P), H) = \int_{s \in CH(P)} d^2(s, H) ds$. Let H_{pca} be the plane going through the center of gravity, parallel to the side $a_{pca} \times b_{pca}$ of $BB_{pca(3,3)}(P)$, and H_{opt} be the bisector of $BB_{opt(P)}$ parallel to the side $a_{opt} \times b_{opt}$. By Lemma 1, part ii) H_{pca} is the best fitting plane of P and therefore,

$$d^2(CH(P), H_{pca}) \leq d^2(CH(P), H_{opt}). \tag{23}$$

We obtain an estimation for γ by determining a lower bound on $d^2(CH(P), H_{pca})$ that depends on c_{pca}, and an upper bound on $d^2(CH(P), H_{opt})$ that depends on c_{opt}. Having an arbitrary bounding box of $CH(P)$ (with side lengths a, b, and c, $a \geq b \geq c$), we denote by H_{ab} the plane going through the center of gravity, parallel to the side $a \times b$. The volume of $CH(P)$ can be expressed as

$$V \quad = \quad V(CH(P)) = \int_0^c \int_0^b \int_0^a \chi_{CH(P)}(x,y,z)dxdydz = \int_0^c g(z)dz,$$

where $\chi_{CH(P)}(x,y,z)$ is the *characteristic function* of $CH(P)$ defined as

$$\chi_{CH(P)}(x,y,z) = \begin{cases} 1 & (x,y,z) \in CH(P) \\ 0 & (x,y,z) \notin CH(P), \end{cases}$$

and $g(z) = \int_0^b \int_0^a \chi_{CH(P)}(x,y,z)dxdy$ is the area of the intersection of $CH(P)$ with the horizontal plane at height z. As before we call $g(z)$ the density function of $CH(P)$. Let c_1 denote the z-coordinate of the center of gravity of $CH(P)$. The line l_{c_1} ($y = c_1$) divides the volume of $CH(P)$ into V_1 and V_2 (see Fig. 13 (a) for an illustration).

Note that $g(z)$ is continuous, but in general not convex in the interval $[0, b]$. Therefore, we cannot use linear functions to derive a lower and an upper bound on the function $d^2(CH(P), H_{ab})$, as we did in Subsection 4.2, because a linear function can intersect $g(z)$ more than once, and we cannot apply Theorem 6. We will show that instead of linear functions, quadratic functions can be used.

Proposition 3 *Let $g(z)$ be the density function of $CH(P)$ defined as above, and let $f(z) = kz^2$ be the parabola such that $\int_0^{c_1} f(z)dz = \int_0^{c_1} g(z)dz$. Then, $\exists c_0 \in [0, c_1]$ such that $f(z) \leq g(x)$ for all $z \leq c_0$ and $f(z) \geq g(z)$ for all $z \geq c_0$.*

Proof. We give a constructive proof. Let $c_0 := \inf\{\ d\ |\forall z \in [d, c_1]\ g(z) \leq f(z)\}$. If $c_0 = 0$, then $f(z) = g(z)$, and the proposition holds. If $c_0 > 0$, then consider the polygon which is the intersection of $CH(P)$ with the plane $z = c_0$. We fix a point p_0 in $CH(P)$ with z-coordinate 0 and construct a pyramid Q by extending all rays from p_0 through the polygon up to the plane $z = c_1$ (see Fig. 12 for an illustration). Since, $f(c_0) = g(c_0)$ the quadratic function $f(z)$ is the density function of Q. Therefore, since the part of Q below c_0 is completely included in $CH(P)$, we can conclude that $f(z) \leq g(z)$ for all $z \leq c_0$. On the other hand, $f(z) \geq g(x)$ for all $z \geq c_0$ by the definition of c_0. □

Now, we present a lower and an upper bound on the variance $d^2(CH(P), H_{ab})$, from which we can derive a bound on $\gamma = \frac{c_{pca}}{c_{opt}}$.

Lemma 11 *The variance $d^2(CH(P), H_{ab})$ is bounded from below by $\frac{7}{256}Vc^2$.*

Proof. We split the integral $\int_0^c (z - c_1)^2 g(z)dz$ at c_1, and prove upper bounds on both parts in the following way: For the left part consider the parabola $f_1(z) = \frac{h_1}{c_1^2}z^2$ such that $\int_0^{c_1} f_1(z)dz = \int_0^{c_1} g(z)dz = V_1$ (see Fig. 13 (b) for an illustration). From $\int_0^{c_1} f_1(z)dz = V_1$ we have that $f_1(z) = \frac{3V_1}{c_1^3}z^2$. Since $f_1(z)$ and $g(z)$ define the same volume on the interval

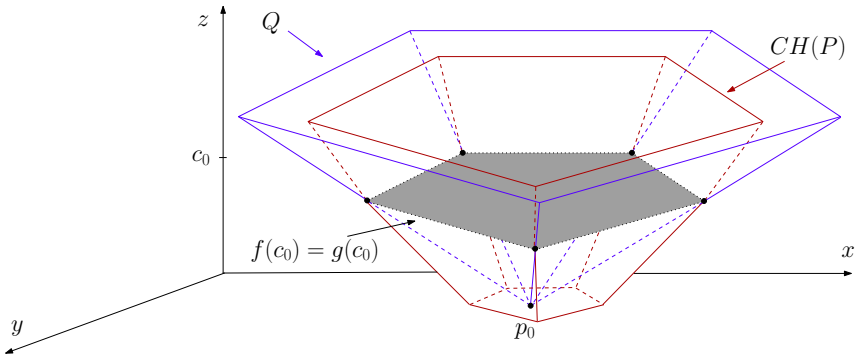


Figure 12. Construction of the intersection of $f(z)$ and $g(z)$.

$[0, c_1]$, they must intersect, and by Proposition 3 we know that if $f_1(z) \neq g(z)$, then they can intersect only once, at a point $c' \in (0, c_1)$. Under these conditions, we can apply Theorem 6, and obtain

$$\int_0^{c_1}(z-c_1)^2 g(z)dz \geq \int_0^{c_1}(z-c_1)^2 f_1(z)dz = \int_0^{c_1}(z-c_1)^2 \tfrac{3V_1}{c_1^3} z^2 dz = \tfrac{V_1 c_1^2}{10}. \quad (24)$$

Analogously, for the right part consider the parabola $f_2(z) = \frac{h_2}{(c_1-c)^2}(z-c)^2 = \frac{h_2}{c_2^2}(z-c)^2$ such that $\int_{c_1}^{c} f_2(y)dy = \int_{c_1}^{c} g(z)dz = V_2$ (see Fig. 13 (b) for an illustration). From $\int_{c_1}^{c} f_2(y)dy = V_2$ we have that $f_1(z) = \frac{3V_2}{c_2^3}(z-c)^2$. By similar arguments as above in the

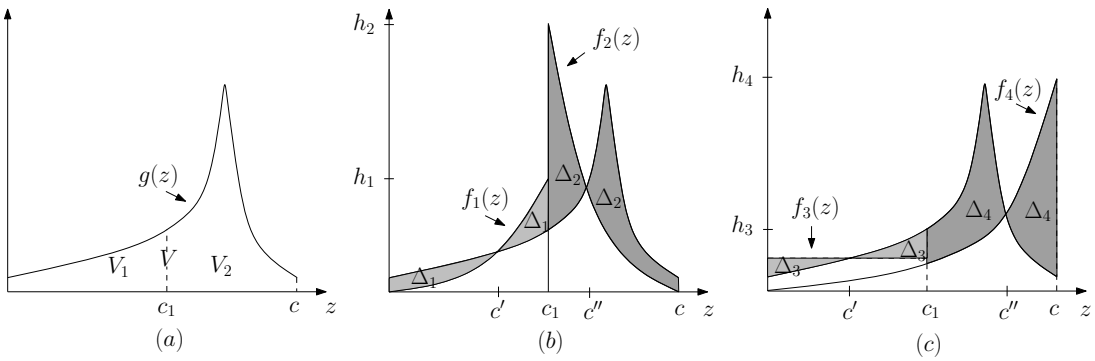


Figure 13. Construction of the lower and upper bounds on $d^2(CH(P), H_{ab})$

case of $f_1(z)$, we can show that $g(z)$ and $f_2(z)$ intersect only once, at a point $c'' \in (c_1, c)$. Applying Theorem 6 we have that

$$\begin{aligned} \int_{c_1}^{c}(z-c_1)^2 g(z)dz &\geq \int_{c_1}^{c}(z-c_1)^2 f_2(z)dz = \int_{c_1}^{c}(z-c_1)^2 \tfrac{3V_2}{c_2^3}(z-c)^2 dz \\ &= \tfrac{V_2 c_2^2}{10}. \end{aligned} \quad (25)$$

From (24) and (25) we obtain that

$$d^2(CH(P), H_{ab}) = \int_0^{c_1}(z-c_1)^2 g(z)dz + \int_{c_1}^{c}(z-c_1)^2 g(z)dz \geq \frac{V_1 c_1^2}{10} + \frac{V_2 c_2^2}{10}.$$

From the Grünbaum-Hammer-Mityagin theorem, we know that $V_1, V_2 \in [\frac{27}{64}V, \frac{37}{64}V]$. Also, we know that $c_1, c_2 \in [\frac{1}{4}c, \frac{3}{4}c]$. It is not hard to show that, under these constrains, the expression $\frac{V_1 c_1^2}{10} + \frac{V_2 c_2^2}{10}$ achieves its minimum of $\frac{7}{256}Vc^2$ for $V_1 = \frac{27}{64}V, c_1 = \frac{3}{4}c$ or $V_1 = \frac{37}{64}V, c_1 = \frac{1}{4}c$. □

Lemma 12 *The variance $d^2(CH(P), H_{ab})$ is bounded from above by $\frac{12729}{71680}Vc^2$.*

The proof of Lemma 12 is similar to the proof of Lemma 11. Here, the functions we use to derive the upper bound on $d^2(CH(P), H_{ab})$ are given in Fig 13 (*c*) (functions $f_3(z)$ and $f_4(z)$).

As a consequence of Lemma 11 and Lemma 12, we have the following upper bound on γ.

Proposition 4 $\gamma < 2.5484$.

Proof. By Lemma 11, we have

$$\frac{7}{256}V{c_{pca}}^2 \leq d^2(CH(P), H_{pca}). \tag{26}$$

On the other hand, by Lemma 12, it follows that

$$d^2(CH(P), H_{opt}) \leq \frac{12729}{71680}V{c_{opt}}^2, \tag{27}$$

From (26), (27) and (23), we obtain

$$\gamma = \frac{c_{pca}}{c_{opt}} \leq \sqrt{\frac{12729}{1960}} < 2.5484.$$

□

We are now ready to present a new parametrized bound on $\lambda_{3,3}(P)$, which is good for a large values of η and θ. The additional crucial relation we exploit in its derivation is the fact given in the following lemma.

Lemma 13 *Let $(x_1, x_2, \ldots, x_d)$ and $(y_1, y_2, \ldots, y_d)$ be two sets of orthogonal base vectors in $\mathbb{R}^d$. For any point set $P \in \mathbb{R}^d$ it holds that*

$$\sum_{i=1}^{d} var(P, x_i) = \sum_{i=1}^{d} var(P, y_i).$$

Proof. We have that

$$\sum_{i=1}^{d} \mathrm{var}(P, x_i) = \sum_{i=1}^{d} \frac{1}{n} \sum_{p \in P} d^2(p, H_{x_i}),$$

where H_{x_i} is a hyperplane orthogonal to the vector x_i, passing through the origin of the coordinate system, $d^2(p, H_{x_i})$ denotes the Euclidean distance of p to H_{x_i}, and $n = |P|$. Since $\sum_{i=1}^{d} d^2(p, H_{x_i})$ is the squared distance of p to the origin of the coordinate system, it can be expressed as the sum of squared distances to the $(d-1)$-dimensional hyperplanes spanned by any set of orthogonal base vectors. Therefore,

$$\sum_{i=1}^{d} d^2(p, H_{x_i}) = \sum_{i=1}^{d} d^2(p, H_{y_i}), \quad \text{and}$$

$$\begin{aligned} \textstyle\sum_{i=1}^{d} \mathrm{var}(P, x_i) &= \tfrac{1}{n} \textstyle\sum_{p \in P} \sum_{i=1}^{d} d^2(p, H_{x_i}) = \tfrac{1}{n} \textstyle\sum_{p \in P} \sum_{i=1}^{d} d^2(p, H_{y_i}) \\ &= \textstyle\sum_{i=1}^{d} \mathrm{var}(P, y_i). \end{aligned}$$

When P is a continuous point set,

$$\mathrm{var}(P, x_i) = \frac{1}{\mathrm{Vol}(P)} \int_{p \in P} d^2(p, H_{x_i}) ds$$

and the claim can be shown as in the discrete case. □

Lemma 14 *$\lambda_{3,3}(P) \leq 6.43\sqrt{1 + \frac{1}{\eta^2} + \frac{1}{\theta^2}}$ for any point set P with aspect ratios $\eta(P) = \eta$ and $\theta(P) = \theta$.*

Proof. Let $x_{pca}, y_{pca}, z_{pca}$ be a set of basis vectors that determine the direction of $BB_{pca(3,3)}(P)$, and let $x_{opt}, y_{opt}, z_{opt}$ be a set of basis vectors that determine the direction of $BB_{opt}(CH(P))$. By Lemma 13, we have that

$$\begin{aligned} &\mathrm{var}(CH(P), x_{pca}) + \mathrm{var}(CH(P), y_{pca}) + \mathrm{var}(CH(P), z_{pca}) = \\ &\mathrm{var}(CH(P), x_{opt}) + \mathrm{var}(CH(P), y_{opt}) + \mathrm{var}(CH(P), z_{opt}). \end{aligned} \tag{28}$$

By Lemma 1, part i), the variance of $CH(P)$ in the direction x_{pca} is the biggest possible, and therefore,

$$\mathrm{var}(CH(P), x_{pca}) \geq \mathrm{var}(CH(P), x_{opt}). \tag{29}$$

Combining (28) and (29) we obtain

$$\begin{aligned} &\mathrm{var}(CH(P), y_{pca}) + \mathrm{var}(CH(P), z_{pca}) \leq \\ &\mathrm{var}(CH(P), y_{opt}) + \mathrm{var}(CH(P), z_{opt}). \end{aligned} \tag{30}$$

We denote by $H_{a_pb_p}$ the plane orthogonal to z_{pca}, going through the center of gravity, and parallel with the side $a_{pca}b_{pca}$ of $BB_{pca(3,3)}(P)$. Similarly, we define $H_{a_pc_p}$, $H_{a_ob_o}$ and $H_{a_oc_o}$. We can rewrite (30) as

$$\begin{aligned} &d^2(CH(P), H_{a_pb_p}) + d^2(CH(P), H_{a_pc_p}) \leq \\ &d^2(CH(P), H_{a_ob_o}) + d^2(CH(P), H_{a_oc_o}). \end{aligned} \tag{31}$$

By Lemma 11, the lower bound on $d^2(CH(P), H_{a_pb_p})$ is $\frac{7}{256}Vc_{pca}^2$, and the lower bound on $d^2(CH(P), H_{a_pc_p})$ is $\frac{7}{256}Vb_{pca}^2$. By Lemma 12, the upper bound on $d^2(CH(P), H_{a_ob_o})$ is $\frac{12729}{71680}Vc_{opt}^2$, and the lower bound on $d^2(CH(P), H_{a_oc_o})$ is $\frac{12729}{71680}Vb_{opt}^2$. Plugging these bounds into (31) we obtain

$$\frac{7}{256}Vc_{pca}^2 + \frac{7}{256}Vb_{pca}^2 \leq \frac{12729}{71680}Vc_{opt}^2 + \frac{12729}{71680}Vb_{opt}^2. \tag{32}$$

Applying $\gamma = \frac{c_{pca}}{c_{opt}}$ in (32), we obtain

$$\frac{7}{256}b_{pca}^2 \leq \left(\frac{12729}{71680} - \frac{7}{256}\gamma\right)c_{opt}^2 + \frac{12729}{71680}b_{opt}^2. \tag{33}$$

By Proposition 4, it follows that $\frac{12729}{71680} - \frac{7}{256}\gamma \geq 0$, and since $b_{opt} \geq c_{opt}$, we get from (33) that

$$\beta = \frac{b_{pca}}{b_{opt}} \leq \sqrt{12.99 - \gamma^2}. \tag{34}$$

The expression $\sqrt{12.99 - \gamma^2}\,\gamma$ $(\geq \beta\,\gamma)$ has its maximum of 6.495 for $\gamma \approx 2.5484$. This together with the bound $\alpha \leq \sqrt{1 + \frac{1}{\eta^2} + \frac{1}{\theta^2}}$ gives

$$\lambda_{3,3}(P) = \alpha\,\beta\,\gamma \leq 6.495\sqrt{1 + \frac{1}{\eta^2} + \frac{1}{\theta^2}}.$$

□

Lemma 10 gives us a bound on $\lambda_{3,3}(P)$ which is good for small values of η and θ. In contrary, the bound from Lemma 14 behaves worse for small values of η and θ, but better for big values of η and θ. Therefore, we combine both of them to obtain the final upper bound.

Theorem 9 *The PCA bounding box of a point set P in $\mathbb{R}^3$ computed over $CH(P)$ has a guaranteed approximation factor $\lambda_{3,3} < 7.81$.*

Proof. The theorem follows from the combination of the two parametrized bounds from Lemma 10 and Lemma 14:

$$\lambda_{3,3} \leq \sup_{\eta\geq 1,\,\theta\geq 1}\left\{\min\left(\eta\,\theta\left(1 + \tfrac{1}{\eta^2} + \tfrac{1}{\theta^2}\right)^{\frac{3}{2}},\ \ 6.495\sqrt{1 + \tfrac{1}{\eta^2} + \tfrac{1}{\theta^2}}\ \right)\right\}.$$

By numerical verification we obtained that the supremum occurs at ≈ 7.807. □

5. Open Problems

Improving the upper bound on $\lambda_{3,3}$, $\lambda_{2,2}$ and $\lambda_{2,1}$, as well as obtaining an upper bound on $\lambda_{3,2}$ is of interest. The approaches we exploit to obtain the upper bounds require an estimation of the length ratios between all corresponding side pairs of the minimum-volume bounding box and the PCA bounding box. However, even in $\mathbb{R}^4$, we do not know how to obtain the estimations of the length ratios for all corresponding side pairs. We believe that obtaining upper bounds on the approximation factor on the quality of PCA bounding boxes in arbitrary dimension requires different approaches than those presented in this paper.

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*, 15:387–396, 1996.
2. G. Barequet and S. Har-Peled. Efficiently approximating the minimum-volume bounding box of a point set in three dimensions. *J. Algorithms*, 38(1):91–109, 2001.
3. N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger. The R^*-tree: An efficient and robust access method for points and rectangles. *ACM SIGMOD Int. Conf. on Manag. of Data*, pages 322–331, 1990.
4. D. Dimitrov, C. Knauer, K. Kriegel, and G. Rote. New upper bounds on the quality of the PCA bounding boxes in R^2 and R^3. In *Proc. 23rd ACM Symp. on Computational Geometry*, 2007.
5. D. Dimitrov, C. Knauer, K. Kriegel, and G. Rote. Upper and lower bounds on the quality of the PCA bounding boxes. In *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision - WSCG 2007*, pages 185–192, 2007.
6. S. Gottschalk, M. C. Lin, and D. Manocha. OBBTree: A hierarchical structure for rapid interference detection. *Computer Graphics*, 30:171–180, 1996.
7. B. Grünbaum. Partitions of mass-distributions and convex bodies by hyperplanes. *Pacific J. Math.*, 10:1257–1261, 1960.
8. I. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 2nd ed., 2002.
9. H. Kharaghani and B. Tayfeh-Rezaie. A Hadamard matrix of order 428. In *J. Combin. Designs 13*, pages 435–440, 2005.
10. M. Lahanas, T. Kemmerer, N. Milickovic, D. B. K. Karouzakis, and N. Zamboglou. Optimized bounding boxes for three-dimensional treatment planning in brachytherapy. In *Med. Phys. 27*, pages 2333–2342, 2000.
11. B. Mityagin. Two inequalities for volumes of convex bodies. *Math. Notes*, 5:61–65, 1968.
12. J. O'Rourke. Finding minimal enclosing boxes. In *Int. J. Comp. Info. Sci. 14*, pages 183–199, 1985.
13. N. Roussopoulos and D. Leifker. Direct spatial search on pictorial databases using packed R-trees. In *ACM SIGMOD*, pages 17–31, 1985.
14. T. Sellis, N. Roussopoulos, and C. Faloutsos. The R^+-tree: A dynamic index for multidimensional objects. In *VLDB Journal*, pages 507–518, 1987.

15. G. Toussaint. Solving geometric problems with the rotating calipers. In *IEEE MELECON*, May 1983.
16. D. V. Vranić, D. Saupe, and J. Richter. Tools for 3D-object retrieval: Karhunen-Loeve transform and spherical harmonics. In *IEEE 2001 Workshop Multimedia Signal Processing*, pages 293–298, 2001.