

Lossless Shape Representation using Invariant Statistics: the Case of Point-sets

Mireille Boutin, Kiryung Lee and Mary Comer
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907 USA
{mboutin,lee505,comerm}@ecn.purdue.edu

Abstract—Boutin and Kemper have shown that the set of unlabeled pairwise distances between the points of a generic point-set in \mathbb{R}^n is a lossless representation of the shape of the point-set. In this paper, we extend this result to the case where each of the points observed is drawn from a similar spherical Gaussian distribution in \mathbb{R}^2 . More precisely, we consider the distribution of the (squared) distance between two points independently drawn from a mixture of spherical Gaussians, each Gaussian having the same variance σ^2 . We then show that two generic such mixtures of spherical Gaussians have the same shape (i.e., they are related by a rigid motion) if and only if their distribution of distances are the same.

I. INTRODUCTION

The motivating application for this paper is the problem of browsing through a large database of shapes. This is a difficult problem on which a considerable effort is currently being expended [1], [2]. It is generally acknowledged that one of the keys to efficiently solving this problem is to have a good representation for the shape of an object. Many objects can be adequately represented by a set of distinguished points called *landmarks* [3]. In many cases, these landmarks are indistinguishable, so they cannot be robustly labeled. The problem of recognizing such objects thus boils down to recognizing the shape of a point-set without labeling. However, in practice, the points observed are noisy. One is thus interested in computing the probability that two observed (unlabeled) point-sets come from two point distributions which have the same shape. In the following, we show that, under some assumptions, two point distributions have the same shape if and only if their distribution of distances (to be defined below) is the same. Thus, the distribution of distances of a point-set distribution is a lossless representation of its shape. This means that the probability that the underlying point distributions have the same shape is equal to the probability that their distribution of distances is the same.

II. THE DETERMINISTIC CASE

Consider a planar point-set $p_1, \dots, p_n \in \mathbb{R}^2$. Denote by Δ_{ij} the square of the Euclidean distance between p_i and p_j ,

$$\Delta_{ij} = \|p_i - p_j\|^2.$$

We put all these square distances in a bag, forgetting their label. In other words, we consider the unordered set of all Δ_{ij} 's of the point set p_1, \dots, p_n ,

$$\{\Delta_{ij}\}_{i \neq j},$$

which we call the *bag of distances* of the point-set. Kemper and Boutin [4] have proposed to use the bag of distances of a point-set to represent its *shape*, i.e. to represent the point-set up to a global rotation, reflection and translation. It is obvious that two point-sets which have the same shape also have the same bag of distances. However, having the same bag of distances does not necessarily imply having the same shape. A counter-example is given by the point-sets

$$(0, 0), (4, 0), (3, 1), (-3, 1) \text{ and} \\ (0, 0), (4, 0), (1, -1), (3, -1),$$

which have different shapes but have the same set of pairwise distances:

$$\{\sqrt{2}, \sqrt{2}, 2, \sqrt{10}, \sqrt{10}, 4\}.$$

However, Boutin and Kemper have shown that such counter-examples are very rare because they lie on a measure zero set. More precisely, they have shown that the following holds.

Theorem 1: [5] There exists a polynomial f in $2n$ variables such that if the points $p_1, p_2, \dots, p_n \in \mathbb{R}^2$ satisfy $f(p_1, p_2, \dots, p_n) \neq 0$, then for any other point-set $\bar{p}_1, \bar{p}_2, \dots, \bar{p}_n$ having the same bag of distances as that of p_1, p_2, \dots, p_n , there exists an orthogonal matrix $M \in \mathbb{R}^{2 \times 2}$ and a translation vector $T \in \mathbb{R}^2$ such that

$$\bar{p}_i = Mp_i + T, \text{ for all } i = 1, \dots, n.$$

Thus the vast majority of point-sets, which we call *generic* point-sets, do not share their bag of distances with any other point-set with a different shape.

In [5], Boutin and Kemper conducted numerical experiments to estimate the likelihood of encountering a non-generic point-set when picking the points on a fixed grid. Although it was observed that non-generic point sets are potentially quite likely on a small grid, the results of their experiments suggest that they are almost never encountered when the points are specified by 15 random digits. The bag of distances thus appears to be a very good representation for the shape of a

point-set, even when the points coordinates are specified by floating point values.

III. GENERALIZATION TO THE NON-DETERMINISTIC CASE

The question remains as to how to use the bag of distances to compare the shape of point-sets that are close, but not exactly equal. It would be tempting to assume that when two bags of distances are *close* in some sense, then the shape of the underlying point-sets are also close, whenever the latter are generic. Unfortunately, this is not true. Indeed, one can pick, respectively, two generic point-sets near two non-generic point-sets having the same bag of distances. The bag of distances of the generic point-sets can be made arbitrarily close by choosing the points closer and closer to their non-generic neighbor. However, their underlying shapes remain different.

The right way to approach the non-deterministic case appears to be the following. We consider the pdf $\rho_i(x)$ of the point \mathbf{p}_i (now considered to be a bivariate random variable whose components are independent), where $x \in \mathbb{R}^2$. We assume that $\rho_i(x)$ is a spherical Gaussian distribution with mean μ_i and variance σ^2 , so that

$$\rho_i(x) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}\|x-\mu_i\|^2}. \quad (1)$$

We then consider the *point-set density* $\rho(x)$ given by the Gaussian mixture

$$\rho(x) = \sum_{i=1}^n \frac{1}{n} \rho_i(x) = \sum_{i=1}^n \frac{1}{n} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}\|x-\mu_i\|^2}. \quad (2)$$

We say that such a point-set density is *generic* if the means of each Gaussian form a generic point-set, i.e. if $f(\mu_1, \mu_2, \dots, \mu_n) \neq 0$ where f is the polynomial function of Theorem 1.

Consider two random variables $\mathbf{p}_1, \mathbf{p}_2$ chosen independently at random according to $\rho(x)$. The square of the distance between these two points is a random variable Δ . Denote its pdf by $r(\Delta)$. We call $r(\Delta)$ the distribution of distances of the point-set density $\rho(x)$. Obviously, the distribution of distances of two point-set densities which have the same shape is the same. The converse statement also holds, under some mild assumptions. Indeed, we have the following theorem, whose proof is given in the last section of this paper.

Theorem 2: Suppose that two generic mixtures of n spherical Gaussians $\rho(x), \bar{\rho}(x)$ have the same variance σ^2 . Then $\rho(x)$ and $\bar{\rho}(x)$ have the same distribution of distances if and only if there exists an orthogonal matrix $M \in \mathbb{R}^{2 \times 2}$ and a translation vector $T \in \mathbb{R}^2$ such that

$$\rho(x) = \bar{\rho}(Mx + T).$$

In other words, under these assumptions, $r(\Delta)$ is a lossless representation of the shape of $\rho(x)$.

IV. NUMERICAL TESTS

Consider two generic point-set densities. Given samples from their respective distributions of distances, one can use the Kolmogorov-Smirnov *goodness of fit* test to estimate

the probability p that the two sets of distance samples come from the same underlying distribution. By Theorem 2, the probability that the underlying point-set densities have the same shape is equal to p .

One of the strengths of the Kolmogorov-Smirnov test is that it applies regardless of the type of underlying distribution. However, the true probability p is an asymptotic value; the estimate obtained with a limited number of samples may or may not be close to p . It is thus interesting to test the accuracy of the Kolmogorov-Smirnov test on finite sets of distance samples coming from either similar or different point-set densities and to check how accurate the results actually are. Overall, we found that the test very accurately identifies distance samples coming from point-set densities with different shapes. However, a large number of distance samples had to be used. In particular, simply using the set of pairwise distances between the means of each Gaussian did not yield a reliable estimate; we had to use many more. Moreover, we found that the test did not reliably identify samples coming from point-sets with the same shape. However, the average over several repeated tests, using different samples, seemed to be a good indicator, as it oscillated around one half whenever the point-set distributions were the same, and quickly dropped to zero as we moved the point-sets far away from each other.

Figure 1 illustrates the results of some of the experiments we performed. In these particular experiments, we used Matlab to generate random point-sets in a unit square and used these point-sets as the means of spherical Gaussian mixtures. For Shape 1, we used a point-set containing ten points. Using the Matlab `unidrnd` and `randn` functions, we generated pairs of iid random points from a mixture of spherical Gaussians with variance $\sigma^2 = 0.04$. We thus generated two sets of 200 distance samples. These two sets of distance samples were then compared with the Kolmogorov-Smirnov test. The result was supposed to be an estimate for the probability that the two sets of samples come from the same distribution. However, we did not obtain a value close to one, as we would have expected. So we repeated the experiment a total of 1000 times, each time with different distance samples. The results are displayed in a histogram in the upper right corner of the figure. The distribution of the results appears to be more or less uniformly distributed on the interval $[0, 1]$. The mean value, at 0.5073, is very close to one half, which is consistent with the results we obtained with all our other simulations.

We also performed several experiments where we constructed point-sets with different shapes and compared large number of their distance samples using the Kolmogorov-Smirnov test. For example, Shape 4 was randomly generated inside a unit square and compared to Shape 1. We generated 200 distance samples from the distribution of distances of Shape 4 and compared them to those of Shape 1. The results were consistently very small (with a mean of 0.0053), as one would expect. We then went on to generate shapes similar to Shape 1 by adding Gaussian noise to each of the points of Shape 1. For example, Shape 2 and Shape 3 were generated using standard deviations equal to 0.1 and 0.4, respectively.

Means of Gaussian Mixture

Histogram of test results

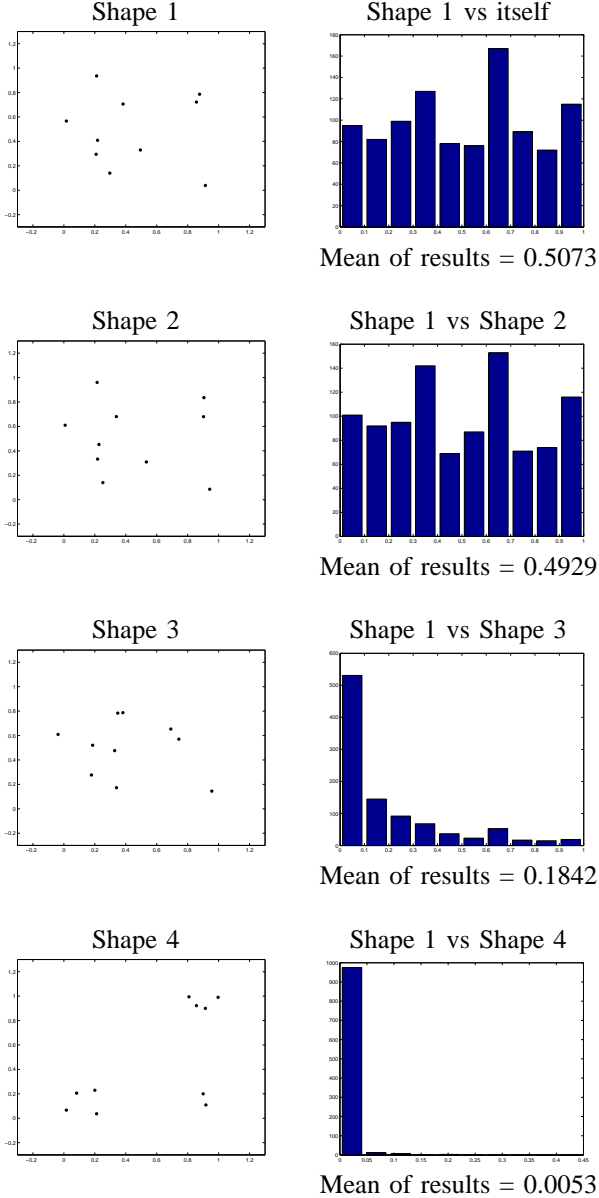


Fig. 1. Comparing samples of the distribution of distances of point-sets with the Kolmogorov-Smirnov test. Shape 1 and Shape 4 are randomly generated shapes which are quite different. Shape 2 is a small random perturbation of Shape 1 while Shape 3 is another, more different, random perturbation of Shape 1. Although a single KS test is unreliable, the value of the mean of the results over several repeated tests clearly indicates whether the underlying point-set distributions are the same.

What we observe is that, as more noise is added to the points, the mean of the KS tests gets smaller and smaller and the distribution of the results peaks more and more near zero. Indeed, the mean of the results displayed in the figure are 0.4929 and 0.1842 respectively, the smaller number corresponding to the more different shape.

V. PROOF OF THEOREM 2

Let

$$\begin{aligned} \rho(x) &= \sum_{i=1}^n \frac{1}{n} \rho_i(x) \\ &= \sum_{i=1}^n \frac{1}{n} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2} \|x - \mu_i\|^2} \end{aligned}$$

be a mixture of 2D spherical Gaussians, all with the same variance σ^2 . Assume that $\mu_1, \mu_2, \dots, \mu_n$ is a generic point-set in \mathbb{R}^2 . We consider the distribution of the random variable Δ defined as the square of the distance between two random variables $\mathbf{p}_1, \mathbf{p}_2 \in \mathbb{R}^2$ chosen independently according to $\rho(x)$.

The distribution $r(\Delta)$ remains unchanged under a rigid motion of the point set density $\rho(x)$ because the random variable $\Delta = \mathbf{p}_1 - \mathbf{p}_2$ is invariant under a simultaneous rigid transformation of the values of \mathbf{p}_1 and \mathbf{p}_2 . Thus the "if" of Theorem 2 is clear. To show the "only if", we need to show that the coordinates of the means μ_i 's of the Gaussian mixture can be determined from $r(\Delta)$.

Claim 1: The moment generating function of Δ is

$$M_{\Delta}(t) = \frac{1}{n} \frac{1}{1 - 4\sigma^2 t} + \frac{2}{n^2} \sum_{i=1}^n \sum_{j=i+1}^n \frac{e^{-\frac{\|\mu_i - \mu_j\|^2 t}{1 - 4\sigma^2 t}}}{1 - 4\sigma^2 t}$$

Proof: By definition, we have

$$\begin{aligned} M_{\Delta}(t) &= E(e^{t\Delta}), \\ &= E\left(e^{t\|\mathbf{x}_1 - \mathbf{x}_2\|^2}\right), \\ &= \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} e^{t\|\mathbf{x}_1 - \mathbf{x}_2\|^2} \rho(\mathbf{x}_1) \rho(\mathbf{x}_2) d\mathbf{x}_1 d\mathbf{x}_2, \\ &= \sum_{i,j=1}^n \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} e^{t\|\mathbf{x}_1 - \mathbf{x}_2\|^2} \frac{\rho_i(\mathbf{x}_1)}{n} \frac{\rho_j(\mathbf{x}_2)}{n} d\mathbf{x}_1 d\mathbf{x}_2. \end{aligned}$$

Let us consider each term of the sum separately. We set

$$M_{ij}(t) := \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} e^{t\|\mathbf{x}_1 - \mathbf{x}_2\|^2} \rho_i(\mathbf{x}_1) \rho_j(\mathbf{x}_2) d\mathbf{x}_1 d\mathbf{x}_2.$$

The quantity $M_{ij}(t)$ represents the expected value of $e^{t\|\mathbf{x}_1 - \mathbf{x}_2\|^2}$ when $\mathbf{x}_1 \sim N(\mu_i, \sigma^2 I)$ and $\mathbf{x}_2 \sim N(\mu_j, \sigma^2 I)$. Writing

$$(\mathbf{X}, \mathbf{Y}) := \mathbf{x}_1 - \mathbf{x}_2,$$

we have $(\mathbf{X}, \mathbf{Y}) \sim N(\mu_i - \mu_j, 2\sigma^2 I)$. Since \mathbf{X} and \mathbf{Y} are independent, we have

$$M_{ij}(y) = E\left(e^{t(\mathbf{X}^2 + \mathbf{Y}^2)}\right) = E\left(e^{t\mathbf{X}^2}\right) E\left(e^{t\mathbf{Y}^2}\right) \quad (3)$$

Let $(\mu_x, \mu_y) = \mu_i - \mu_j$.

$$\begin{aligned}
E(e^{t\mathbf{X}^2}) &= \int_{-\infty}^{\infty} \frac{e^{tx^2} e^{-\frac{(x-\mu_x)^2}{4\sigma^2}}}{\sqrt{2\pi}\sqrt{2}\sigma} dx \\
&= \int_{-\infty}^{\infty} \frac{e^{\left(tx^2 - \frac{(x-\mu_x)^2}{4\sigma^2}\right)}}{\sqrt{2\pi}\sqrt{2}\sigma} dx \\
&= \int_{-\infty}^{\infty} \frac{e^{\left(-\frac{1-4\sigma^2 t}{4\sigma^2} \left(x - \frac{\mu_x}{1-4\sigma^2 t}\right)^2 + \frac{t\mu_x^2}{1-4\sigma^2 t}\right)}}{2\sqrt{\pi}\sigma} dx \\
&= \frac{e^{\frac{t\mu_x^2}{1-4\sigma^2 t}}}{\sqrt{1-4\sigma^2 t}} \int_{-\infty}^{\infty} \frac{e^{-\frac{\left(x - \frac{\mu_x}{1-4\sigma^2 t}\right)^2}{2(\sqrt{2}\sigma/\sqrt{1-4\sigma^2 t})^2}}}{\frac{2\sqrt{\pi}\sigma}{\sqrt{1-4\sigma^2 t}}} dx \\
&= \frac{1}{\sqrt{1-4\sigma^2 t}} e^{\frac{t\mu_x^2}{1-4\sigma^2 t}} \quad (4)
\end{aligned}$$

By symmetry, we also have

$$E(e^{t\mathbf{Y}^2}) = \frac{1}{\sqrt{1-4\sigma^2 t}} e^{\frac{t\mu_y^2}{1-4\sigma^2 t}}. \quad (5)$$

Combining Equation 3 with Equations 4 and 5, we get

$$M_{ij}(t) = \frac{1}{1-4\sigma^2 t} e^{\frac{\|\mu_i - \mu_j\|^2 t}{1-4\sigma^2 t}}$$

Finally, we get

$$\begin{aligned}
M_{\Delta}(t) &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n M_{ij}(t) \\
&= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \frac{1}{1-4\sigma^2 t} e^{\frac{\|\mu_i - \mu_j\|^2 t}{1-4\sigma^2 t}} \\
&= \frac{1}{n} \frac{1}{1-4\sigma^2 t} + \\
&\quad + \frac{2}{n^2} \sum_{i=1}^n \sum_{j=i+1}^n \frac{1}{1-4\sigma^2 t} e^{\frac{\|\mu_i - \mu_j\|^2 t}{1-4\sigma^2 t}}
\end{aligned}$$

To simplify the notation, we relabel the pairwise distances between the means as

$$\begin{aligned}
d_1 &= \|\mu_1 - \mu_2\|^2, \\
d_2 &= \|\mu_1 - \mu_3\|^2, \\
&\vdots \\
d_N &= \|\mu_{n-1} - \mu_n\|^2,
\end{aligned}$$

where $N = \frac{n(n-1)}{2}$. The moment generating function of Δ then becomes

$$M_{\Delta}(t) = \frac{1}{n} \frac{1}{1-4\sigma^2 t} + \frac{2}{n^2} \sum_{i=1}^N \frac{e^{\frac{d_i t}{1-4\sigma^2 t}}}{1-4\sigma^2 t}. \quad (6)$$

Claim 2: The k th order moment $E(\Delta^k)$ can be written as

$$E(\Delta^k) = \sum_{r=0}^k F_r(\sigma^2) \Pi_r(d_1, \dots, d_N),$$

where F_r is some function, and Π_r is the power sum

$$\Pi_r(d_1, \dots, d_N) = \sum_{1 \leq i \leq N} d_i^r.$$

Proof: We begin by obtaining the power series expansion for the moment generating function $M_{\Delta}(t)$. Using the power series for the exponential and the binomial formula, we have

$$\begin{aligned}
M_{\Delta}(t) &= \frac{1}{n^2} \sum_{i=1}^N \frac{1}{1-4\sigma^2 t} e^{\frac{d_i t}{1-4\sigma^2 t}}, \\
&= \frac{1}{n^2} \sum_{i=1}^N \frac{1}{1-4\sigma^2 t} \sum_{m=0}^{\infty} \frac{1}{m!} \frac{(d_i t)^m}{(1-4\sigma^2 t)^m}, \\
&= \frac{1}{n^2} \sum_{m=0}^{\infty} \sum_{i=1}^N \frac{1}{m!} \frac{(d_i t)^m}{(1-4\sigma^2 t)^{m+1}}, \\
&= \frac{1}{n^2} \sum_{m=0}^{\infty} \sum_{i=1}^N \frac{1}{m!} (d_i t)^m \sum_{r=0}^{\infty} c_{m,r} (4\sigma^2 t)^r, \\
&= \frac{1}{n^2} \sum_{m,r=0}^{\infty} \frac{c_{m,r}}{m!} (4\sigma^2)^r \left(\sum_{i=1}^N d_i^m \right) t^{m+r},
\end{aligned}$$

where $c_{m,r}$ are the coefficients of the binomial expansion. The conclusion follows from the fact that the k th order moment is equal to $k!$ times the coefficient of t^k in the above series. \blacksquare

Claim 3: If two point-set densities

$$\begin{aligned}
\rho(x) &= \sum_{i=1}^n \frac{1}{n} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2} \|x - \mu_i\|^2} \\
\bar{\rho}(x) &= \sum_{i=1}^n \frac{1}{n} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2} \|x - \bar{\mu}_i\|^2}
\end{aligned}$$

with the same σ^2 have the same distribution of distances $r(\Delta) = \bar{r}(\Delta)$, then the means of their respective Gaussians have the same bag of distances.

Proof: If the distributions of distances are the same, then all the moments of the distances are the same for both distributions. By Claim 2, and since the variance σ^2 is the same for both distribution, this means that the power sums are the same for both distributions.

$$\Pi_r(d_1, \dots, d_N) = \Pi_r(\bar{d}_1, \dots, \bar{d}_N), \text{ for all } r.$$

This means that the values of the first N elementary symmetric functions

$$\begin{aligned}
S_1(d_1, \dots, d_N) &= \sum_{1 \leq i \leq N} d_i, \\
S_2(d_1, \dots, d_N) &= \sum_{1 \leq i < j \leq N} d_i d_j, \\
S_3(d_1, \dots, d_N) &= \sum_{1 \leq i < j < k \leq N} d_i d_j d_k, \\
&\vdots \\
S_N(d_1, \dots, d_N) &= \prod_{1 \leq i \leq N} d_i,
\end{aligned}$$

are also the same for both distributions:

$$S_r(d_1, \dots, d_N) = S_r(\bar{d}_1, \dots, \bar{d}_N) \text{ for all } r = 1, \dots, N.$$

(The explicit relationship between S_p and Π_1, \dots, Π_p is given by the so-called Newton-Girard formulas.) We then consider the following equations,

$$x^n - \sum_{k=1}^N \Pi_k x^{N-k} = 0,$$

whose solutions set is $\{d_1, \dots, d_N\} = \{\bar{d}_1, \dots, \bar{d}_N\}$, i.e. the bag of distances of the point-set defined by the means of both Gaussian mixtures, which thus must be the same, as claimed. ■

Now suppose that two mixtures of spherical Gaussians with the same σ are such that they have the same distribution of distances. Then, by Claim 3, the point-set formed by their

means must have the same bag of distances, and thus, by Theorem 1, the Gaussian mixtures must be the same up to a rigid motion.

REFERENCES

- [1] P. Shilane, P. Min, M. M. Kazhdan, and T. A. Funkhouser, "The princeton shape benchmark." in *SMI*. IEEE Computer Society, 2004, pp. 167–178.
- [2] D. Bespalov, C. Y. Ip, W. C. Regli, and J. Shaffer, "Benchmarking cad search techniques." in *Symposium on Solid and Physical Modeling*, L. Kobbelt and V. Shapiro, Eds. ACM, 2005, pp. 275–286.
- [3] L. D. F. Costa and R. M. C. Jr., *Shape Analysis and Classification: Theory and Practice*. CRC, 2000.
- [4] M. Boutin and G. Kemper, "On reconstructing n -point configurations from the distribution of distances or areas," *Adv. Appl. Math.*, vol. 32, pp. 709–735, 2004.
- [5] —, "Which point configurations are determined by the distribution of their pairwise distances," *Int. J. Compt. Geometry and Appl.*, in press.