A fast and efficient method for solving the multiple generalized circle detection problem

Rudolf Scitovski
Department of Mathematics, University of Osijek

In collaboration with:

Kristian Sabo, Department of Mathematics, University of Osijek

Budapest, October 6-8, 2018

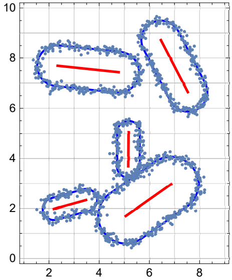
This work was supported by the Croatian Science Foundation through research grants:
The optimization and statistical models and methods in recognizing properties of data sets measured with errors (OSMoMeSIP-IP-2016-06-6545) and
Methodological framework for efficient energy management by intelligent data analytics IP-2016-06-8350

Problem statement

$$\mathcal{A} = \{a^i = (x_i, y_i) \in \Delta \subset \mathbb{R}^2 \colon \Delta = [\alpha_1, \beta_1] \times [\alpha_2, \beta_2],\ i = 1, \dots, m\}$$

(set of data points coming from a number of generalized circles in the plane not known in advance that should be reconstructed or detected)

Assumption: Number of data points coming from a generalized circle $\mathcal{C}$ depends on $|\mathcal{C}|$



$$\mathcal{C}(\mu, \nu, r) = \{x \in \mathbb{R}^2 \colon D(\mathrm{seg}(\mu, \nu), x) = r\},$$
$$\mathrm{seg}(\mu, \nu) = \{x \in \mathbb{R}^2 \colon x = \lambda\mu + (1 - \lambda)\nu,\ \lambda \in [0, 1]\},$$

Applications – may arise in research in various areas like biology, medicine, etc.

Problem statement - center-based clustering

$\Pi(\mathcal{A}) = \{\pi_1, \ldots, \pi_k\}$, $\mathcal{A} \subset \mathbb{R}^n$ (k-partition)

$$\bigcup_{i=1}^{k} \pi_i = \mathcal{A}, \quad \pi_r \cap \pi_s = \emptyset, \quad r \neq s, \quad |\pi_j| \geq 1, \qquad j = 1, \ldots, k$$

$d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, $d(u, v) = \|u - v\|_2^2$ (LS-distance-like function)

$c_j := \underset{x \in \operatorname{conv}(\pi_j)}{\operatorname{argmin}} \sum_{a^i \in \pi_j} d(x, a^i)$ (centroid of the cluster π_j)

Global optimization problem (GOP)

$$\underset{\Pi \in \mathcal{P}(\mathcal{A};k)}{\operatorname{argmin}} \mathcal{F}(\Pi), \qquad \mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a^i \in \pi_j} d(c_j, a^i)$$

$\mathcal{P}(\mathcal{A}; k)$ (set of all partitions)

Problem statement - center-based clustering

Conversely, let $z_1, \ldots, z_k \in \mathbb{R}^n$ be the set of points

Minimal distance principle gives the partition $\Pi = \{\pi(z_1), \ldots, \pi(z_k)\}$

$$\pi(z_j) = \{a \in \mathcal{A} : d(z_j, a) \leq d(z_s, a), \forall s = 1, \ldots, k\}, \qquad j = 1, \ldots, k.$$

Global optimization problem:

$$\operatorname*{argmin}_{z \in \Delta^k} F(z), \quad F(z) = \sum_{i=1}^{m} \min_{1 \leq j \leq k} d(z_j, a^i) \quad z = (z_1, \ldots, z_k).$$

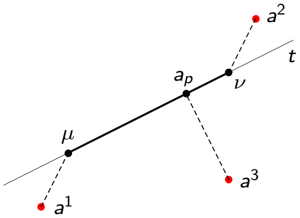
In the case when cluster-centers are generalized circles ($\mathcal{C}$-cluster centers)

$$\operatorname*{argmin}_{\mu,\nu \in \Delta^k,\, r \in (0,R]^k} F(\mu, \nu, r), \quad F(\mu, \nu, r) = \sum_{i=1}^{m} \min_{1 \leq j \leq k} \mathfrak{D}(\mathcal{C}_j(\mu_j, \nu_j, r_j), a^i),$$

$R = \frac{1}{2} \min\{\alpha_2 - \alpha_1, \beta_2 - \beta_1\}$,

$\mathfrak{D}(\mathcal{C}_j, a^i)$ (distance from the point a^i to the generalized circle $\mathcal{C}_j$)

Problem statement - distance $\mathfrak{D}$ from the point to the generalized circle



$$D(\mathrm{seg}(\mu,\nu),a) = \min_{t\in\mathrm{seg}[\mu,\nu]} \|t-a\|_2 = \begin{cases} \|a-\mu\|_2, & (\nu-\mu)^T(a-\mu) \le 0,\\ \|a-\nu\|_2, & (\mu-\nu)^T(a-\nu) < 0,\\ \left\| \mu\frac{(\nu-\mu)^T(\nu-a)}{\|\mu-\nu\|_2^2} + \left(1-\frac{(\nu-\mu)^T(\nu-a)}{\|\mu-\nu\|_2^2}\right)\nu \right\|_2, & \text{else.} \end{cases}$$

$$\mathfrak{D}_1(\mathcal{C}(\mu,\nu,r),a) = |D(\mathrm{seg}(\mu,\nu),a) - r| \quad \text{(LAD-distance)}$$

$$\mathfrak{D}_2(\mathcal{C}(\mu,\nu,r),a) = (D(\mathrm{seg}(\mu,\nu),a) - r)^2 \quad \text{(TLS-distance)}$$

$$\mathfrak{D}(\mathcal{C}(\mu,\nu,r),a) = (D(\mathrm{seg}(\mu,\nu),a)^2 - r^2)^2 \quad \text{(algebraic distance)}$$

The method

Multiple generalized circle detection problem:

$$\underset{\mu,\nu\in\Delta^k,\, r\in(0,R]^k}{\operatorname{argmin}} F(\mu,\nu,r), \quad F(\mu,\nu,r) = \sum_{i=1}^{m} \min_{1\le j\le k} \mathfrak{D}(\mathcal{C}_j(\mu_j,\nu_j,r_j),a^i)$$

Possibilities:

(i) Application of global optimization algorithm `DIRECT` [1-2];

(ii) Modification of k-means algorithm for the case of $\mathcal{C}$-cluster-centers [3-4];

- Our proposal: (ii) with a very carefully chosen initial approximation

[1] D.R.Jones, C.D.Perttunen, B.E.Stuckman, *Journal of Optimization Theory and Applications* **79**(1993)

[2] D.E.Finkel, `DIRECT` Optimization Algorithm User Guide (2003), `http://www4.ncsu.edu/~ctk/Finkel_Direct/DirectUserGuide_pdf.pdf`

[3] T.Marošević, R.Scitovski, *Croatian Operational Research Review* **6**(2015)

[4] R.Grbić, D.Grahovac, R.Scitovski, *Pattern Recognition* **60**(2016)

Algorithm 1

Step A: (Assignment step) For each set of mutually different generalized circles $\mathcal{C}_1(\mu_1, \nu_1, r_1), \ldots, \mathcal{C}_k(\mu_k, \nu_k, r_k)$, the data set $\mathcal{A}$ should be partitioned into k disjoint nonempty clusters $\pi_1, \ldots, \pi_k$ by using the minimal distance principle for $j = 1, \ldots, k$

$$\pi_j := \{a \in \mathcal{A} \colon \mathfrak{D}(\mathcal{C}_j(\mu_j, \nu_j, r_j), a) \leq \mathfrak{D}(\mathcal{C}_s(\mu_s, \nu_s, r_s), a),\ s \neq j\};$$

Step B: (Update step) For the given partition $\Pi = \{\pi_1, \ldots, \pi_k\}$ of the set $\mathcal{A}$, we should determine the corresponding $\mathcal{C}$-cluster-centers $\hat{\mathcal{C}}_j(\hat{\mu}_j, \hat{\nu}_j, \hat{r}_j)$, $j = 1, \ldots, k$ by solving the following GOPs:

$$\underset{\mu,\nu\in\Delta,\ r\in[0,R]}{\operatorname{argmin}} \sum_{a\in\pi_j} \mathfrak{D}(\mathcal{C}(\mu, \nu, r), a), \quad j = 1, \ldots, k.$$

Stopping criterion according to A.M.Bagirov, :
$\frac{F_{k-1}-F_k}{F_1} < \epsilon_B, \quad F_k = F(\mu^{(k)}, \nu^{(k)}, r^{(k)}), \quad \epsilon_B > 0$

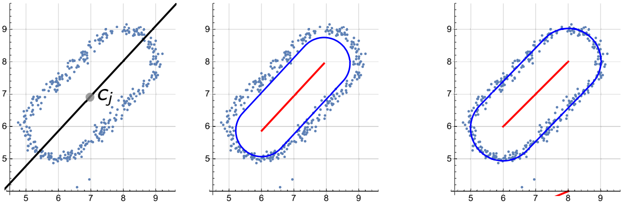
[1] A.M.Bagirov, Pattern Recognition, 2008

$$\underset{\mu,\nu\in\Delta,\, r\in[0,R]}{\operatorname{argmin}} \sum_{a\in\pi_j} \mathfrak{D}(\mathcal{C}(\mu,\nu,r),a)$$

Estimation

- of the length of the line-segment (see [1]): $\delta_j = \frac{2}{|\pi_j|} \sum_{s\in\pi_j} |(a^i - c_j)\cdot v_1^{(j)}|$;
- of the line-segment $\operatorname{seg}(\mu_j,\nu_j) \approx [c_j - \frac{1}{2}\delta_j v_1^{(j)}, c_j + \frac{1}{2}\delta_j v_1^{(j)}]$;
- of the radius of the generalized circle $\mathcal{C}_j$: $r_j \approx \frac{1}{|\pi_j|} \sum_{s\in\pi_j} |(a^i - c_j)\cdot v_2^{(j)}|$;

DIRECT-domain: $K_{\mu_j} \times K_{\nu_j} \times K_{r_j}$, where $K_u = \{x\in\mathbb{R}^2 : \|u-x\|_\infty \le \frac{1}{4}\delta_j\}$, $u\in\{\mu_j,\nu_j\}$ and $K_{r_j} = \{x\in\mathbb{R} : |x-r_j| \le \frac{1}{4}\delta_j\}$



[1] J.C.R.Thomas, Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 2011

The proposed method

Application of `KMCC` algorithm

- with a good enough initial partition (DBSCAN);
- with a good enough approximation of $\mathcal{C}$-cluster-centers (Incremental algorithm for line detection)

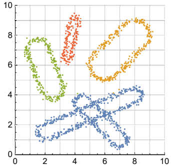
Application of DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm with parameters $MinPts > 2$, and $\epsilon > 0$

Estimation of the parameter ϵ

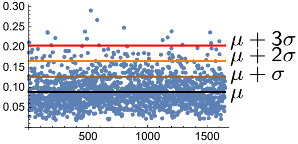
ϵ_a, $a \in \mathcal{A}$ - radius of the circle containing $MinPts > 2$ elements from $\mathcal{A}$

$\mathcal{R} = \{\epsilon_a \colon a \in \mathcal{A}\}$

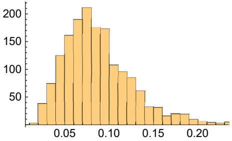
$$\epsilon = \mu + 3\sigma, \quad \mu = \mathrm{Mean}(\mathcal{R}), \quad \sigma^2 = \frac{1}{m-1} \sum_{a \in \mathcal{A}} (\epsilon_a - \mu)^2$$



(a) Sequence (ϵ_a)



(b) Histogram of the sequence (ϵ_a)



[1] M.Ester and H.P.Krieogel, J.Sander, *2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 1996

[2] P.Viswanath, V.S.Babu, *Pattern Recognition Letters* **30**(2009)

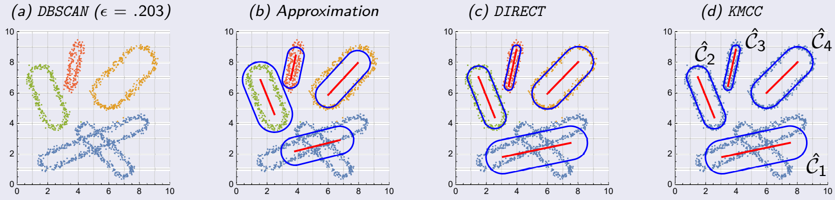
[3] Y.Zhu, K.M.Ting, M.J.Carman, *Pattern Recognition* **60**(2016)

Algorithm 3 (KMCC by using a good initial partition)

Input: $\mathcal{A} \subset \mathbb{R}^2$; *MinPts*;
1: Determine parameter ϵ in the previously described way;
2: By using the DBSCAN algorithm determine the initial partition $\Pi^0 = \{\pi_1, \ldots, \pi_\kappa\}$;
3: By using the KMCC algorithm and the initial partition Π^0 determine generalized circles $\mathcal{C}_1^\star, \ldots, \mathcal{C}_\kappa^\star$;

Output: $\{\mathcal{C}_1^\star, \ldots, \mathcal{C}_\kappa^\star\}$.

Example 1



QAD=(0.831,0.139,0.174,0.155)

The question is, when does a calculated circle detect some generalized circle, and when not?

- $\mathcal{A} \subset \mathbb{R}^2$ - a set of data points coming from a number of generalized circles $\mathfrak{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ not known in advance
- $\Pi = \{\hat{\pi}_1, \ldots, \hat{\pi}_\kappa\}$ - an optimal κ-partition of the set $\mathcal{A}$
- $\hat{\mathcal{C}}_1, \ldots, \hat{\mathcal{C}}_\kappa$ - $\mathcal{C}$-cluster-centers
- $\epsilon > 0$ - parameter from `DBSCAN-algorithm`

$V_j = \{\mathfrak{D}_1(\hat{\mathcal{C}}_j, a) \colon a \in \hat{\pi}_j\}$ - set of LAD-distances

$\mathtt{QAD[j]} := \mu_j + 2\sigma_j, \ \mu_j = \mathrm{Mean}\,(V_j), \ \sigma_j^2 = \frac{1}{|V_j|-1} \sum\limits_{v_j \in V_j} (v_j - \mu_j)^2$ - quantile absolute deviation

Due to the homogeneity property of the set $\mathcal{A}$, if $\mathtt{QAD[j]} < \epsilon$, then at least 95% of points of the cluster $\hat{\pi}_j$ have a smaller or equal LAD-distances from the circle $\hat{\mathcal{C}}_j$

If for some $j_0 \in \{1, \ldots, \kappa\}$, $\mathtt{QAD}[j_0] > \epsilon$, the generalized circle $\hat{\mathcal{C}}_{j_0}$ has not been recognized as one of generalized circles from $\mathfrak{C}$

Density-Based Clustering index (DBC):

$$\mathtt{DBC}(\kappa) := \max_{j=1,\ldots,\kappa} \mathtt{QAD[j]}$$

Example

Let as assume that among the original generalized circles $\mathcal{C}_1, \ldots, \mathcal{C}_k$ there are some that touch or intersect.

Algorithm 4 (KMCC by using a good initial $\mathcal{C}$-cluster-centers)

Input: $\mathcal{A} \subset \mathbb{R}^2$;

According to (Nievergelt, 1994), determine the best TLS-line of the set $\mathcal{A}$;

Solve GOP $\underset{\alpha\in[-\pi/2,\pi/2],\delta\in[0,\beta_2]}{\operatorname{argmin}} \Phi(\alpha,\delta)$, $\Phi(\alpha,\delta) = \sum_{i=1}^{m} \min\{d_1^{(i)}, \mathfrak{D}_H(\ell(\alpha,\delta), a^i)\}$, where $d_1^{(i)} = \frac{(u_1x_i+v_1y_i+z_1)^2}{u_1^2+v_1^2}$, and $\mathfrak{D}_H(\ell(\alpha,\delta), a^i) = (x_i\cos\alpha + y_i\sin\alpha - \delta)^2$ and $\ell(\alpha,\delta) \equiv x\cos\alpha + y\sin\alpha - \delta = 0$ is line in the Hesse normal form. Denote the solution by $\mathcal{L}_2$;

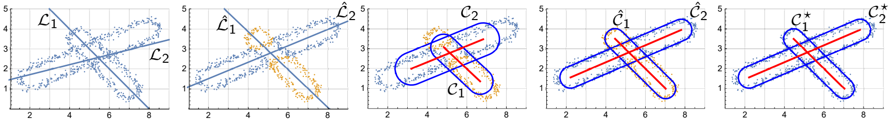
By using KMCL-algorithm determine lines $\hat{\mathcal{L}}_1, \hat{\mathcal{L}}_2$;

By using Algorithm 2 determine the corresponding $\mathcal{C}$-cluster-centers $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2$;

By using the KMCC algorithm and initial centers $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2$ determine generalized circles $\mathcal{C}_1^\star, \mathcal{C}_2^\star$;

Output: $\{\mathcal{C}_1^\star, \mathcal{C}_2^\star\}$.

(a) Incremental alg. (b) KMCL algorithm (c) Application of Algorithm 2 (e) KMCC algorithm



Numerical experiments

Let us first choose integer $k \in \{2, 3, 4, 5\}$ and real numbers $\gamma = 2.25$ and $\delta = 1.5$. Choose line-segments $\mathrm{seg}(\mu_j, \nu_j)$, $j = 1, \ldots, k$, such that $\mu_j, \nu_j \in [1.2, 8.8]^2 \subset \Delta$, $\Delta \subset [0, 10] \times [0, 10]$ and that the following holds:

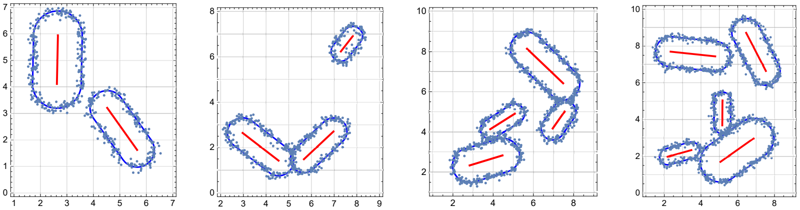
(i) $\|\mu_j - \nu_j\| < \gamma$,

(ii) mutual distances between midpoints of line-segments are greater than δ,

(iii) distances between vertices of each segment and other segments are greater than δ.

(iv) radii of generalized circles $r_j \in \mathcal{U}(0.25, 1.0)$.

On the generalized circle $\mathcal{C}_j$, depending on its length, we choose $m_j = \lceil 30|\mathcal{C}_j|\rceil$ uniformly distributed points and then to each point in the direction of the normal vector we add a random error from normal distribution with expectation 0 and the variance $\sigma^2 = 0.1$.



Result analysis for generalized circles

No. of circles	Generalized circles detected							CPU-time (DIRECT)	(KMCC)	Total
	0	1	2	3	4	5	6			
2	0	0	29	0	0	0	0	.123	.102	0.225
3	0	0	0	30	2	0	0	.175	.156	0.331
4	0	0	0	1	26	0	0	.253	.295	0.548
5	0	0	1	0	0	11	1	.322	.311	0.633

Construction of data in the neighborhood of a generalized circle

$m = 30|\mathcal{C}|$ - number of uniformly distributed random points in the neighbourhood of the generalized circle $\mathcal{C}(\mathrm{seg}(\mu, \nu), r)$

- Random points in the neighbourhood of the circle $C(S_0, r)$, $S_0 = \frac{1}{2}(\mu + \nu)$ are translated into the left and the right edge of the generalized circle $\mathcal{C}$;
- Random points in the neighbourhood of line-segments $\overline{AB}$, $\overline{CD}$
  $T_i = \lambda_i A + (1 - \lambda_i)B + \epsilon_i, \quad A = \mu - n_0 r,\ B = \nu - n_0 r$
  $\lambda_i \in \mathcal{U}(0, 1), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2 I)$