Directional Statistics of Preferential Orientations of Two Shapes in Their Aggregate and Its Application to Nanoparticle Aggregation

March 10, 2017

Abstract

Nanoscientists have long conjectured that adjacent nanoparticles aggregate with one another in certain preferential directions during a chemical synthesis of nanoparticles, which is referred to the oriented attachment. For the study of the oriented attachment, the microscopy and nanoscience communities have used dynamic electron microscopy for direct observations of nanoparticle aggregation and have been so far relying on manual and qualitative analysis of the observations. We propose a statistical approach for studying the oriented attachment quantitatively with multiple aggregation examples in imagery observations. We abstract an aggregation by an event of two primary geometric objects merging into a secondary geometric object. We use a point set representation to describe the geometric features of the primary objects and the secondary object, and formulated the alignment of two point sets to one point set to estimate the orientation angles of the primary objects in the secondary object. The estimated angles are used as data to estimate the probability distribution of the orientation angles and test important scientific hypotheses statistically. The general approach was applied for our motivating example, which demonstrated that nanoparticles of certain geometries have indeed preferential orientations in their aggregates.

Keywords: Point-set-based shape representation, Shape alignment, Orientation of shapes, Statistical analysis of circular data

1 INTRODUCTION

A particle aggregation is a merging of two smaller particles into one larger particle, which is one of the main driving forces that grow atoms or molecular clusters into nanoparticles during a chemical synthesis of nanoparticles. With a better understanding of a particle aggregation, synthesizing nanoparticles of desired sizes and shapes should be possible (Welch et al. 2016, Zhang et al. 2012, Li et al. 2012).

As seen in Figure 1, a particle aggregation is essentially a two-step process, a collision of two primary particles followed by their restructuring to a larger secondary particle. Some collisions effectively lead to a subsequent restructuring (or coalescence), while other collisions are ineffective. The degree of effectiveness depends on how primary nanoparticles are spatially oriented in a collision. When primary particles are oriented ineffectively, they become separate again or rotate to a preferred orientation, as in the phenomenon known as *the oriented attachment* (Li et al. 2012). A fundamental scientific problem to solve is to study the oriented attachment, which can be achieved by directly observing and analyzing a number of nanoparticle aggregation cases. This paper addresses how to study the microscopic observations of nanoparticle aggregations to statistically analyze the preferential orientations of primary nanoparticles.

A major contribution of this paper is to provide a mathematical foundation for statistically studying the oriented attachment. The microscopy and nanoscience communities have been relying on manual analysis of a very few examples of nanoparticle aggregation for the study of the oriented attachment. Our proposed method will provide a systematic way of statistically analyzing a large population of aggregation examples to find a statistically

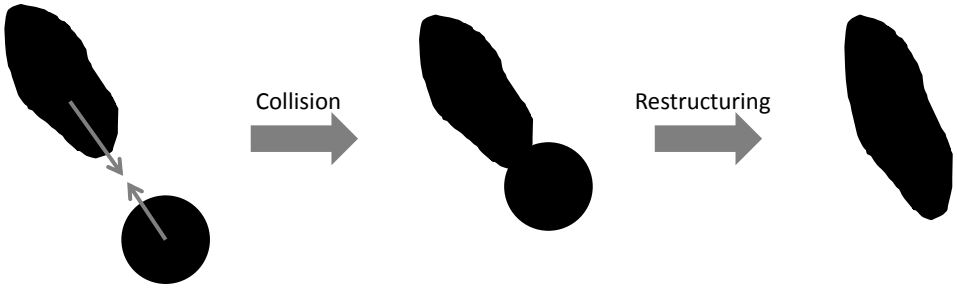


Figure 1: Particle aggregation

reliable estimate of the preferential orientations of nanoparticles within their aggregations. We acknowledge that there are some existing works on the statistical analysis of aggregates in concrete and asphalt engineering (Mora & Kwan 2000, Wang 1999), but those works primarily focused on studying how aggregates are sized and shaped, instead of studying how aggregating components are oriented. We believe that our work is the first of its kind in statistically studying the oriented attachment.

In addition to the contribution in applications and modeling, this paper contains two methodological contributions. In statistical shape analysis, a problem of aligning one shape to another shape has been well studied to possibly find the relative orientation of one to another (Schmidler 2007, Green & Mardia 2006). However, the existing theory and methods do not work for analyzing the orientations of two aggregating components within their aggregate, which involves aligning two shapes to one shape that is assumed to be a union of the two shapes. This paper presents in Section 3.2 a solution for this two-to-one alignment problem. On the other hand, in directional statistics, angular data and their distribution have long been studied (Fisher 1995), but studies on the probability distribution of angular data with some symmetries are lacking. For our motivating example, the angular distribution of a particle orientation is essentially four-fold symmetric due to geometrical symmetries of nanoparticles. Section 4 presents a new probability distribution to model such symmetries and the related statistical analysis.

The remainder of this paper is organized as follows. Section 2 describes microscopy data that motivated this study. Section 3 presents how we mathematically model an aggregate and the orientations of aggregating components. Section 4 describes several statistical inference problems on the orientations, including a probability density estimation problem and some statistical hypothesis testing problems, which were applied in Section 5 to test several scientific hypotheses posed to explain the oriented attachment. Section 6 provides our conclusions.

2 DATASET

We used dynamic scanning transmission electron microscopy to synthesize and directly observe growth of silver nanoparticles (Woehl et al. 2012), taking a sequence of electron

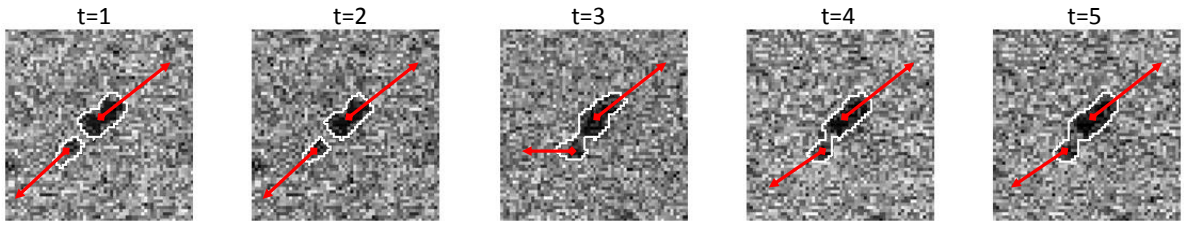


Figure 2: Dynamic microscopy data of particle aggregation

microscope images of about two hundred silver nanoparticles and their aggregations. We applied an object-tracking algorithm (Park et al. 2015) with the microscope images to track their aggregations, which identified 184 different aggregation cases. Figure 2 displays an example of the captured aggregation events.

For each aggregation event, we take two items of information: the first is the image of two primary nanoparticles taken immediately before the aggregation, e.g., the image at $t = 2$ in Figure 2, and the second is the image of the secondary nanoparticle taken immediately after the final aggregation, e.g. the image at $t = 4$. After the final aggregation, the orientations of the two primary nanoparticles do not change due to strong physical forces as shown in Figure 2. Therefore, the aggregate image can be taken any time after the final aggregation, but our choice is the time immediate after the aggregation because the aggregate might later undergo a significant restructuring. The time resolution of the imaging process is faster than a normal aggregation speed, so the 'immediate before the aggregation' and the 'immediate after the aggregation' are well defined from the observed image sequences.

Each of the before images and the after images is two-dimensional, depicting the projection of the three dimensional geometries of nanoparticles on a two-dimensional space. Since the nanoparticles imaged are constrained to a very thin layer of a sample chamber, we assume that geometrical information along the z-direction is relatively insignificant. A set of the image pairs for the 184 aggregation events will be analyzed for studying how primary nanoparticles are oriented in their aggregates.

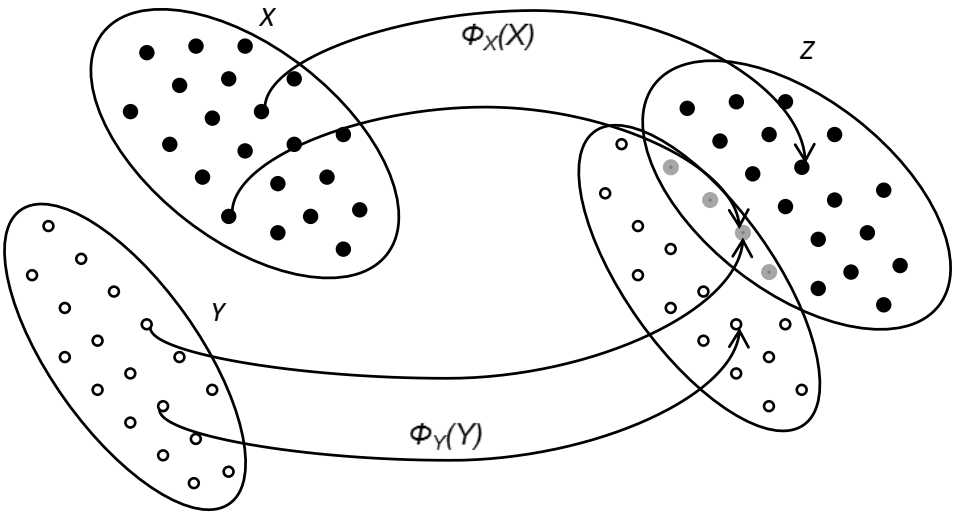


Figure 3: n-covering representation of particle aggregate

3 MODELING AGGREGATION

We abstract an aggregation as a merge of two geometric objects. We first describe how we model geometric objects. Let $\mathbb{X}$ denote a set of all image pixel coordinates in an $H \times W$ digital image,

$$\mathbb{X} := \{(h, w) : h = 0, 1, 2, ..., H, w = 0, 1, 2, ..., W\}.$$

A geometric object imaged on $\mathbb{X}$ is represented by a simply connected subset of $\mathbb{X}$ that represents a set of all image pixel coordinates locating inside the geometric object. The set-based representation has been popularly used for shape analysis (Mémoli & Sapiro 2005, Mémoli 2007), which seems more useful for our motivating problem than other popular shape representation models such as the representation by landmark points (Kendall 1984, Dryden & Mardia 1998) and the representation by a closed curve (Younes 1998, Srivastava et al. 2011). The landmark-based approach has a major technical issue regarding how to manually choose the landmarks of many geometrical bodies, which are also subject to human bias. More importantly, an aggregation of two geometric objects is better represented by the set-based representation. An aggregation of two objects can be naturally represented by the union of two subsets representing the two objects.

Geometric objects move and rotate before they aggregate. The movement and rotation

operations in $\mathbb{X}$ are represented by a Euclidean rigid body transformation. Let $E(\mathbb{X})$ denote a collection of all Euclidean rigid body transformations defined on $\mathbb{X}$. An element ϕ in $\mathbb{X}$ is an Euclidean rigid body transformation that basically shifts $\boldsymbol{x} \in \mathbb{X}$ by $\boldsymbol{c} \in \mathbb{X}$ in the negative direction and rotates the shifting result about the origin by $\theta \in [0, 2\pi]$,

$$\phi(\boldsymbol{x}) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} (\boldsymbol{x} - \boldsymbol{c}). \tag{1}$$

For a set $X \subset \mathbb{X}$, we use a notation $\phi(X)$ to denote the image of X transformed by ϕ,

$$\phi(X) = \{\phi(\boldsymbol{x}); \boldsymbol{x} \in X, \phi \in E(\mathbb{X})\}.$$

When X represents a geometric object, $\phi(X)$ represents the image of the geometric object transformed by the movement and rotation operations defined by ϕ. The operations do not deform the geometric object but just change its poses, i.e., locations and orientations, which is why ϕ is called a rigid body transformation.

Let $X \subset \mathbb{X}$ and $Y \subset \mathbb{X}$ denote two simply connected subsets of $\mathbb{X}$ that represent two primary objects, and let $Z \subset \mathbb{X}$ denote a simply connected subset of $\mathbb{X}$ that represents the aggregate of the two primary objects. The two primary objects may move and rotate before they collide and aggregate. Let $\phi_X \in E(\mathbb{X})$ and $\phi_Y \in E(\mathbb{X})$ denote the Euclidean rigid body transformations that represent the movements and rotations of X and Y before they aggregate. As shown in Figure 3, before the aggregate Z is fully restructured to a different shape, Z is approximately an overlapping union of $\phi_X(X)$ and $\phi_Y(Y)$,

$$Z = \phi_X(X) \cup \phi_Y(Y), \text{ where } \phi_X \in E(\mathbb{X}), \phi_Y \in E(\mathbb{X}).$$

In practice, $\mathbb{X}$ is a digital image, so the equality does not exactly hold due to digitization errors. The aggregate Z can be partitioned into three pieces, $Z_1 = \phi_X(X) \backslash \phi_Y(Y)$, $Z_2 = \phi_Y(Y) \backslash \phi_X(X)$ and $Z_3 = \phi_X(X) \cap \phi_Y(Y)$, where $\backslash$ is a set difference operator. We call the center of mass of Z_3 as an aggregation center, which we denote by $\boldsymbol{c}_{X,Y}$. As we illustrated in Figure 4, we define the orientation of X in Z as the orientation of $\boldsymbol{c}_{X,Y}$ in the standard coordinate system of $\phi_X(X)$. The standard coordinate system for X is defined as a map $T_X : \mathbb{X} \to \mathbb{R}^2$ that assigns a point $x \in X$ to a unique coordinate number $T_X(x)$, which induces the standard coordinate system for $\phi_X(X)$ that maps a point $y \in \phi_X(X)$ to

$T_X \circ \phi_X^{-1}(y)$. Therefore, the orientation of X in Z is

$$\boldsymbol{v}_X = \frac{T_X \circ \phi_X^{-1}(\boldsymbol{c}_{X,Y})}{||T_X \circ \phi_X^{-1}(\boldsymbol{c}_{X,Y})||}, \text{ or } \theta_X = \text{angle}(\boldsymbol{v}_X),$$

where angle($\boldsymbol{v}_X$) is the angular part of the polar coordinate of $\boldsymbol{v}_X$. Similarly, the orientation of Y in Z is defined by

$$\boldsymbol{v}_Y = \frac{T_Y \circ \phi_Y^{-1}(\boldsymbol{c}_{X,Y})}{||T_Y \circ \phi_Y^{-1}(\boldsymbol{c}_{X,Y})||}, \text{ or } \theta_Y = \text{angle}(\boldsymbol{v}_Y).$$

Our primary interest is to study the oriented attachment, i.e., investigating what angles of θ_X and θ_Y are more frequently observed from multiple aggregation examples.

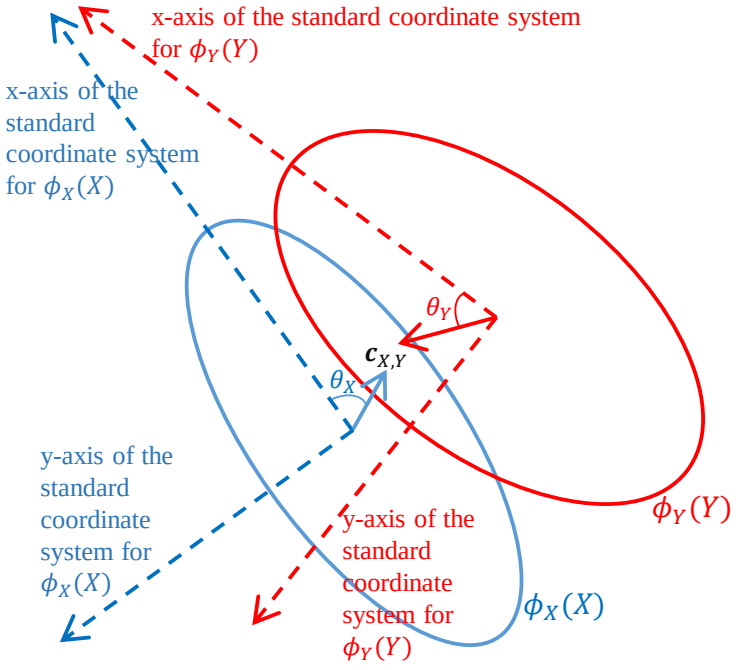


Figure 4: Practical meaning of the definition of θ_X and θ_Y.

The T_X and T_Y are independent of ϕ_X, ϕ_Y and $\boldsymbol{c}_{X,Y}$, i.e., the choice of the former does not affect the latter, and vice versa. Section 3.1 describes how to define and estimate T_X and T_Y, and Section 3.2 describe how to estimate $\phi_X \in E(\mathbb{X})$, $\phi_Y \in E(\mathbb{X})$ and $\boldsymbol{c}_{X,Y}$. Estimating $\phi_X \in E(\mathbb{X})$ implies estimating its parameters $\boldsymbol{c}_X$ and θ_X. Likewise, estimating T_X and T_Y implies estimating the unknown parameters of T_X and T_Y. The parametric forms of T_X and T_Y will be later defined in Section 3.1. The proposed approaches are validated using simulation datasets in Section 3.3.

3.1 Estimation of T_X

The standard coordinate system of X must be consistently defined with those of other geometric objects geometrically similar to X, so their orientations can be defined consistently. To accomplish this, we define a reference shape for a collection of geometric objects geometrically similar to X and define T_X as the Euclidean rigid body transformation that best aligns X to the reference shape. The transformation outcome is invariant to a Euclidean rigid body transformation of X, i.e., $T_X(X) = T_{\phi(X)}(\phi(X))$ for $\phi \in E(\mathbb{X})$, unless the reference shape is redefined, so it provides consistent coordinate numbers for those having similar geometries but different poses. In this section, we describe how we define a reference shape and estimate $T_X \in E(\mathbb{X})$ given a reference shape.

We first work on how to estimate T_X when a reference shape is given. Let X and X_0 denote the simply connected subsets of $\mathbb{X}$ that represent a geometric object and its reference shape respectively. Suppose that X and X_0 consist of m and m_0 point coordinates as follows,

$$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\} \text{ and}$$
$$X_0 = \{\boldsymbol{x}_1^{(0)}, \boldsymbol{x}_2^{(0)}, \ldots, \boldsymbol{x}_{m_0}^{(0)}\},$$

where $\boldsymbol{x}_i \in \mathbb{X}$ denotes the ith element of X, and $\boldsymbol{x}_j^{(0)} \in \mathbb{X}$ indicates the jth element of X_0. We want to find $T_X \in E(\mathbb{X})$ that best aligns X to X_0,

$$T_X(X) \approx X_0,$$

where the closeness of the two sets is measured by a set distance. A popular set distance is the p norm distance (Mémoli 2007), which basically averages the distances between each pair of the elements in the two sets that correspond to each other. Let μ_{ij} define the following measure of correspondence in between the elements of the two sets, $T_X(X)$ and X_0,

$$\mu_{ij} = 1 \text{ if } T_X(\boldsymbol{x}_i) \text{ corresponds to } \boldsymbol{x}_j^{(0)} \text{ and 0 otherwise.} \tag{2}$$

When μ_{ij}'s are known, the set distance is defined by

$$dist(T_X(X), X_0; \boldsymbol{\mu}) = \left(\sum_{i,j} \mu_{ij} \left|\left| \phi(\boldsymbol{x}_i) - \boldsymbol{x}_j^{(0)} \right|\right|^p \right)^{1/p},$$

where $\boldsymbol{\mu}$ denotes a $m \times m_0$ matrix with the (i,j)th element μ_{ij}. The T_X that best aligns X to X_0 can be achieved by minimizing the distance,

$$T_X^*(\boldsymbol{\mu}) = \arg \min_{T_X \in E(\mathbb{X})} dist(T_X(X), X_0; \boldsymbol{\mu}).$$

The expressions for the translation vector $\boldsymbol{c}_{T_X}^*$ and rotation angle $\theta_{T_X}^*$ of $T_X^*(\boldsymbol{\mu})$ can be found at Rangarajan et al. (1997),

$$\begin{aligned} \boldsymbol{c}_{T_X}^* &= \frac{\sum_{i=1}^{m}\sum_{j=1}^{m_0}\mu_{ij}(\boldsymbol{x}_i - \boldsymbol{x}_j^{(0)})}{\sum_{i=1}^{m}\sum_{j=1}^{m_0}\mu_{ij}} \text{ and} \\ \theta_{T_X}^* &= \arctan\left(\frac{\sum_{i=1}^{m}\sum_{j=1}^{m_0}\mu_{ij}(\boldsymbol{x}_j^{(0)} \times \boldsymbol{x}_i)}{\sum_{i=1}^{m}\sum_{j=1}^{m_0}\mu_{ij}(\boldsymbol{x}_j^{(0)} \cdot \boldsymbol{x}_i)}\right), \end{aligned} \tag{3}$$

where $(a_1, a_2) \times (b_1, b_2) = a_1 b_2 - a_2 b_1$ and $(a_1, a_2) \cdot (b_1, b_2) = a_1 b_1 + a_2 b_2$.

However, $\boldsymbol{\mu}$ is unknown. We propose to use the T_X-invariance property of the Euclidean distance matrix of X to estimate $\boldsymbol{\mu}$ so that the estimated $\boldsymbol{\mu}$ can be plugged into equation (3) to estimate T_X. Let us first define the Euclidean distance matrix of X as

$$\boldsymbol{D}(X) = \begin{bmatrix} 0 & d_{\mathbb{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2) & d_{\mathbb{X}}(\boldsymbol{x}_1, \boldsymbol{x}_3) & \dots & d_{\mathbb{X}}(\boldsymbol{x}_1, \boldsymbol{x}_m) \\ d_{\mathbb{X}}(\boldsymbol{x}_2, \boldsymbol{x}_1) & 0 & d_{\mathbb{X}}(\boldsymbol{x}_2, \boldsymbol{x}_3) & \dots & d_{\mathbb{X}}(\boldsymbol{x}_2, \boldsymbol{x}_m) \\ d_{\mathbb{X}}(\boldsymbol{x}_3, \boldsymbol{x}_1) & d_{\mathbb{X}}(\boldsymbol{x}_3, \boldsymbol{x}_2) & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix},$$

where $d_{\mathbb{X}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2$. The distance matrix is invariant under a Euclidean rigid body transformation,

$$\boldsymbol{D}(X) = \boldsymbol{D}(T_X(X)) \text{ for } T_X \in E(\mathbb{X}).$$

In addition, the matrix $\boldsymbol{D}(X)$ contains sufficient information that describes the geometrical features of X, because X is uniquely determined from $\boldsymbol{D}(X)$ up to rotations, reflections and translations by applying the multidimensional scaling to $\boldsymbol{D}(X)$ (Lele 1993, Theorem 1). These two properties allow us to define a T_X-invariant distance between the two geometries, $T_X(X)$ and X_0. Note that X and its reference shape X_0 presumably has similar geometries, so the Euclidean distance matrices of $T_X(X)$ and X_0 should be comparable, i.e.,

$$d_{\mathbb{X}}(T_X(\boldsymbol{x}_i), T_X(\boldsymbol{x}_k)) \approx d_{\mathbb{X}}(\boldsymbol{x}_j^{(0)}, \boldsymbol{x}_l^{(0)}) \text{ for every } \mu_{ij} = 1, \mu_{kl} = 1.$$

Collectively, the equalities are represented by

$$\boldsymbol{D}(T_X(X)) \approx \boldsymbol{\mu}\boldsymbol{D}(X_0)\boldsymbol{\mu}^T.$$

Due to the T_X-invariance of an Euclidean distance matrix, it also implies

$$\boldsymbol{D}(X) \approx \boldsymbol{\mu}\boldsymbol{D}(X_0)\boldsymbol{\mu}^T.$$

Let $d_{\mathbb{D}}(X, X_0; \boldsymbol{\mu}) = ||\boldsymbol{D}(X) - \boldsymbol{\mu}\boldsymbol{D}(X_0)\boldsymbol{\mu}^T||_F$. We will find $\boldsymbol{\mu}$ that minimizes the distance,

$$d_{\mathbb{D}}(X, X_0) = \min_{\boldsymbol{\mu} \in \mathbb{M}_{X,X_0}} d_{\mathbb{D}}(X, X_0; \boldsymbol{\mu}), \tag{4}$$

where $\mathbb{M}_{X,X_0} := \{(\mu_{ij}) \in \{0,1\}^{m \times m_0} : \sum_{i=1}^{m} \mu_{ij} \geq 1, \sum_{j=1}^{m_0} \mu_{ij} \geq 1\}$ defines the range of $\boldsymbol{\mu}$; it was defined to make sure that one element in $T_X(X)$ is mapped to at least one element in X_0 and vice versa. The algorithm to solve the optimization problem in (4) can be found in the online supplementary material. Once $\boldsymbol{\mu}$ is estimated, the estimate can be plugged into equation (3) to estimate the two parameters of T_X. It is noteworthy that there is another way to estimate $\boldsymbol{\mu}$, which finds simultaneously $\boldsymbol{\mu}$ and T_X by solving

$$\min_{\boldsymbol{\mu} \in \mathbb{M}_{X,X_0}} \min_{T_X \in E(\mathbb{X})} dist(T_X(X), X_0; \boldsymbol{\mu}). \tag{5}$$

The optimization has been popularly used for shape matching or two point-set matching (Mémoli 2007). The similar formulation was also proposed in statistical shape analysis (Rangarajan et al. 1997). The optimization is very complicated, because it requires an alternating optimization for T_X and $\boldsymbol{\mu}$ (Rangarajan et al. 1997, Green & Mardia 2006). The alternating procedure often finds local optimality.

Note that $\boldsymbol{D}(X)$ and $\boldsymbol{D}(X_0)$ contain all geometric features of X and X_0 and are also invariant to the Euclidean transformations of X and X_0, so the measure of similarity between the two Euclidean distance matrices (i.e. $d_{\mathbb{D}}(X, X_0)$) can be used as a measure of geometric similarity of X and X_0. Now we use $d_{\mathbb{D}}(X, X_0)$ to group geometric objects by geometric similarities and define a reference shape for each similarity group. Suppose that we have $2N$ primary geometric objects from N different aggregation observations. We first cluster the $2N$ objects into K shape categories. In this paper, we use the k-means clustering with distance $d_{\mathbb{D}}$, where K was chosen using the information criterion, AIC

(Akaike 1992). Suppose that N_k geometric objects are grouped to the kth shape category, and $X_n^{(k)} \subset \mathbb{X}$ denote the nth geometric object from the shape category. We choose a cluster representative of the shape category and define it as a reference shape for the shape category. The cluster representative is chosen among $\{X_n^{(k)}; n = 1, ..., N_k\}$ so that it minimizes the average distance to the other cluster members. If the cluster representative is X_r^k, r should satisfy

$$r = \arg\min_{n=1,...,N_k} \sum_{n'=1}^{N_k} d_{\mathbb{D}}(X_{n'}^{(k)}, X_n^{(k)}).$$

We normalize out the location and orientation of the cluster representative by applying the classical multidimensional scaling to $X_r^{(k)}$. The multidimensional scaling first applies the double centering on $\boldsymbol{D}(X_r^{(k)})$, subsequently takes the eigen-decomposition on the doubly centered matrix, and finally computes the matrix composed of the eigenvectors scaled by the square roots of the corresponding eigenvalues (Lele 1993). Since the rank of $\boldsymbol{D}(X_r^{(k)})$ is two, the output matrix of the multidimensional scaling has two columns, and each row vector of the output matrix represents a point coordinate in $\mathbb{R}^2$. Let $\tilde{X}_r^{(k)}$ denote a set of the row vectors in the matrix. It is easy to verify $\boldsymbol{D}(X_r^{(k)}) = \boldsymbol{D}(\tilde{X}_r^{(k)})$ so $d_{\mathbb{D}}(X_r^{(k)}, \tilde{X}_r^{(k)}) = 0$. Therefore $\tilde{X}_r^{(k)}$ represents the exactly same geometry as $X_r^{(k)}$. The major axis of $\tilde{X}_r^{(k)}$ is always along the x-axis in that the first coordinates of the elements in $\tilde{X}_r^{(k)}$ were generated from the first eigenvector in the multidimensional scaling. Therefore, $\tilde{X}_r^{(k)}$ can be seen as a version of $X_r^{(k)}$ with its orientation normalized. We define $\tilde{X}_r^{(k)}$ as a reference shape for the kth shape category. We will present our simulation study in Section 3.3 for validating the approaches proposed in this section.

3.2 Alignment of primary objects to an aggregate

Let $X \subset \mathbb{X}$ and $Y \subset \mathbb{X}$ denote two primary objects, and let $Z \subset \mathbb{X}$ denote the aggregate of the two primary objects. Suppose that X, Y and Z consist of m_X, m_Y, and m_Z point coordinates respectively,

$$X = \{\boldsymbol{x}_i \in \mathbb{X}; i = 1, \ldots, m_X\}$$
$$Y = \{\boldsymbol{y}_j \in \mathbb{X}; j = 1, \ldots, m_Y\}$$
$$Z = \{\boldsymbol{z}_k \in \mathbb{X}; k = 1, \ldots, m_Z\}.$$

Since $Z = \phi_X(X) \cup \phi_Y(Y)$, some points in Z correspond to the map of X by ϕ_X, and the other points correspond to the map of Y by ϕ_Y. Let $\boldsymbol{\mu}^X = (\mu_{ik}^X)$ denote the point-to-point correspondences from X to Z, and let $\boldsymbol{\mu}^Y = (\mu_{jk}^Y)$ denote the point-to-point correspondence from Y to Z. Please note that the $(\boldsymbol{\mu}^X, \boldsymbol{\mu}^Y)$ ranges for

$$\begin{aligned}\mathbb{M}_{X,Y;Z} = \{(\boldsymbol{\mu}^X, \boldsymbol{\mu}^Y) : &\sum_{k=1}^{m_Z} \mu_{ik}^X \geq 1 \quad \forall i = 1, .., m_X, \\ &\sum_{k=1}^{m_Z} \mu_{jk}^Y \geq 1 \quad \forall j = 1, ..., m_Y, \\ &\sum_{i=1}^{m_X} \mu_{ik}^X + \sum_{j=1}^{m_Y} \mu_{jk}^Y \geq 1 \quad \forall k = 1, ..., m_Z\},\end{aligned}$$

where the first two inequalities imply that each element in X and Y corresponds to at least one element in Z and the last inequality implies that each element in Z corresponds to an element in either X or Y. When $(\boldsymbol{\mu}^X, \boldsymbol{\mu}^Y)$ are known, the two rigid body transformations $\phi_X \in E(2)$ and $\phi_Y \in E(2)$ can be estimated by solving

$$\min_{\phi_X, \phi_Y \in E(2)} \sum_{i=1}^{m_X} \sum_{k=1}^{m_Z} \mu_{i,k}^X \left|\left|\phi_X(\boldsymbol{x}_i) - \boldsymbol{z}_k\right|\right|^2 + \sum_{j=1}^{m_Y} \sum_{k=1}^{m_Y} \mu_{jk}^Y \left|\left|\phi_Y(\boldsymbol{y}_j) - \boldsymbol{z}_k\right|\right|^2.$$

The optimal solution can be obtained by using the first order necessary condition: $\phi_X^* = \boldsymbol{R}(\theta_X^*)(\boldsymbol{x} - \boldsymbol{c}_X^*)$ and $\phi_Y^* = \boldsymbol{R}(\theta_Y^*)(\boldsymbol{x} - \boldsymbol{c}_Y^*)$ with

$$\begin{aligned}\boldsymbol{c}_X^* &= \frac{\sum_{i=1}^{m_X} \sum_{k=1}^{m_Z} \mu_{ik}^X (\boldsymbol{x}_i - \boldsymbol{z}_k)}{\sum_{i=1}^{m_X} \sum_{k=1}^{m_Z} \mu_{ik}^X}, \theta_X^* = \arctan\left(\frac{\sum_{i=1}^{m_X} \sum_{k=1}^{m_Z} \mu_{ik}^X (\boldsymbol{z}_k \times \boldsymbol{x}_i)}{\sum_{i=1}^{m_X} \sum_{k=1}^{m_Z} \mu_{ik}^X (\boldsymbol{z}_k \cdot \boldsymbol{x}_i)}\right) \\ \boldsymbol{c}_Y^* &= \frac{\sum_{j=1}^{m_Y} \sum_{k=1}^{m_Z} \mu_{jk}^Y (\boldsymbol{y}_i - \boldsymbol{z}_k)}{\sum_{j=1}^{m_Y} \sum_{k=1}^{m_Z} \mu_{jk}^Y}, \theta_Y^* = \arctan\left(\frac{\sum_{j=1}^{m_Y} \sum_{k=1}^{m_Z} \mu_{jk}^Y (\boldsymbol{z}_k \times \boldsymbol{y}_j)}{\sum_{j=1}^{m_Y} \sum_{k=1}^{m_Z} \mu_{jk}^Y (\boldsymbol{z}_k \cdot \boldsymbol{y}_j)}\right).\end{aligned} \tag{6}$$

Since $(\boldsymbol{\mu}^X, \boldsymbol{\mu}^Y)$ are unknown, similar to what we did in the previous section, we use the Euclidean distance matrices of X, Y and Z to estimate $(\boldsymbol{\mu}^X, \boldsymbol{\mu}^Y)$,

$$\min_{(\boldsymbol{\mu}^X, \boldsymbol{\mu}^Y) \in \mathbb{M}_{X,Y;Z}} d_{\mathbb{D}}(X, Z; \boldsymbol{\mu}_X) + d_{\mathbb{D}}(Y, Z; \boldsymbol{\mu}_Y). \tag{7}$$

The algorithm to solve the optimization problem can be found in the online supplementary material. The optimal solution provides the point-to-point correspondence $(\boldsymbol{\mu}^X, \boldsymbol{\mu}^Y)$. By plugging $(\boldsymbol{\mu}^X, \boldsymbol{\mu}^Y)$ in the expression (6), the ϕ_X and ϕ_Y can be estimated.

In addition, the aggregation center of Z can be estimated with $(\boldsymbol{\mu}^X, \boldsymbol{\mu}^Y)$ by first finding the subset of Z that corresponds to both X and Y,

$$C_{X,Y} = \{\boldsymbol{z}_k \in Z : \mu_{ik}^X = 1 \text{ and } \mu_{jk}^Y = 1\},$$

and then estimating the mass center of $C_{X,Y}$,

$$\boldsymbol{c}_{X,Y} = \frac{\sum_{\boldsymbol{z}_k \in C_{X,Y}} \boldsymbol{z}_k}{|C_{X,Y}|}, \tag{8}$$

where $|\cdot|$ is the number of elements in a set. This result combine with the estimation of the ϕ_X and ϕ_Y to evaluate $\phi_X^{-1}(\boldsymbol{c}_{X,Y})$ and $\phi_Y^{-1}(\boldsymbol{c}_{X,Y})$.

3.3 Simulation study

We performed a simulation study to numerically validate the proposal approaches described in the previous subsections. We simulated multiple aggregation datasets, where simulation inputs were shape factors of primary objects, the variations of the shape factors, and the levels of observation noises. We restricted the shapes of primary objects to ellipses, for which the shape factors are characterized by the major axis lengths and the minor axis lengths. We followed the following generative procedure to simulate a set of 50 aggregation cases,

Inputs: $\nu_{a,X}$: the logarithm of the mean of the X's major axis length,
$\nu_{a,Y}$: the logarithm of the mean of the Y's major axis length,
$\nu_{b,X}$: the logarithm of the mean of the X's minor axis length,
$\nu_{b,Y}$: the logarithm of the mean of the Y's minor axis length,
σ^2: shape variations, and σ_e^2: noise variance.

Step 1. Simulate X: Sample $\log(a_X) \sim \mathcal{N}(\nu_{a,X}, \sigma^2)$ and $\log(b_X) \sim \mathcal{N}(\nu_{b,X}, \sigma^2)$. Generate a noisy image of an ellipse, $\tilde{X} = \left\{(x_1, x_2) \in \mathbb{X}; \frac{x_1^2}{a_X^2} + \frac{x_2^2}{b_X^2} \leq 1 + \epsilon(|\frac{x_2}{x_1}|)\right\}$, where $\epsilon(|\frac{x_2}{x_1}|) \sim \mathcal{N}(0, \sigma_e^2)$ is a random process depending on $|\frac{x_2}{x_1}|$. Let T_X denote a random Euclidean rigid body transformation with a translation vector $\boldsymbol{c}_{T_X} \sim \text{Uniform}([0, H] \times [0, W])$ and a rotation angle $\theta_{T_X} \sim \text{Uniform}([0, \pi/2])$. The noisy image $\tilde{X}$ is transformed to $T_X^{-1}(\tilde{X})$, which serves X.

Step 2. Simulate Y: Sample $\log(a_Y) \sim \mathcal{N}(\nu_{a,Y}, \sigma^2)$ and $\log(b_Y) \sim \mathcal{N}(\nu_{b,Y}, \sigma^2)$. Generate a noisy image of an ellipse, $\tilde{Y} = \left\{(x_1, x_2) \in \mathbb{X}; \frac{x_1^2}{a_Y^2} + \frac{x_2^2}{b_Y^2} \leq 1 + \epsilon(|\frac{x_2}{x_1}|)\right\}$, where $\epsilon(|\frac{x_2}{x_1}|) \sim \mathcal{N}(0, \sigma_e^2)$ is a random process depending on $|\frac{x_2}{x_1}|$. Let T_Y denote a random Euclidean rigid body transformation with a translation vector $\boldsymbol{c}_{T_Y} \sim \text{Uniform}([0, H] \times [0, W])$ and a rotation angle $\theta_{T_Y} \sim \text{Uniform}([0, \pi/2])$. The noisy image $\tilde{Y}$ is transformed to $T_Y^{-1}(\tilde{Y})$, which serves Y.

Step 3. Simulate Z: Sample $\boldsymbol{c}_X \sim \text{Uniform}(X)$ and $\boldsymbol{c}_Y \sim \text{Uniform}(Y)$. Let ϕ_X denote the Euclidean rigid body transformation with a translation vector $\boldsymbol{c}_X$ and a rotation angle $\theta_X = \pi\text{-angle}(\boldsymbol{c}_X + \boldsymbol{c}_{T_X})$, and let ϕ_Y denote the Euclidean rigid body transformation with a translation vector $\boldsymbol{c}_Y$ and a rotation angle $\theta_Y =\text{-angle}(\boldsymbol{c}_Y + \boldsymbol{c}_{T_Y})$. Let $Z = \phi_X(X) \bigcup \phi_Y(Y)$.

Step 4. Repeat Steps 1 through 3 for 50 times.

We fixed $\nu_{b,X} = \log(5)$ while $\nu_{a,X}$ was varied to $\exp(\nu_{a,X}) = r_X \exp(\nu_{b,X})$, where r_X represents the ratio of the mean major axis length and the mean minor axis length. Similarly, we fixed $\nu_{b,Y} = \log(5)$ and chose $\nu_{a,Y}$ as $\nu_{a,Y} = \log(r_Y) + \nu_{b,Y}$. We fixed $\sigma^2 = 0.03^2$, which makes $\exp(\nu_{b,X})$ or $\exp(\nu_{b,Y})$ approximately range for $[4.5, 5.5]$. We also fixed $\sigma_e^2 = 0.1^2$, which makes $1+\epsilon(|\frac{x_2}{x_1}|)$ approximately range for $[0.97, 1.03]$. We tried six different combinations of $r_X \in \{1.1, 1.4, 2.2\}$ and $r_Y \in \{1.1, 1.4, 2.2\}$ to simulate simulation cases involving different shape factors. For each combination, we have 50 aggregation cases, which serve a simulation dataset.

We applied Sections 3.1 and 3.2 to the six simulated datasets to estimate T_X, T_Y, ϕ_X and ϕ_Y. Note that the T_X is parameterized by two parameters $\boldsymbol{c}_{T_X}$ and θ_{T_X}, T_Y by $\boldsymbol{c}_{T_Y}$ and θ_{T_Y}, ϕ_X by $\boldsymbol{c}_X$ and θ_X, and ϕ_Y by $\boldsymbol{c}_Y$ and θ_Y. The estimated parameters are denoted by $\boldsymbol{c}_{T_X}^*$, $\theta_{T_X}^*$, $\boldsymbol{c}_{T_Y}^*$, $\theta_{T_Y}^*$, $\boldsymbol{c}_X^*$, θ_X^*, $\boldsymbol{c}_Y^*$ and θ_Y^*. For each of the six simulated datasets, we evaluated the differences of the estimated parameter values and the corresponding simulation inputs over 50 simulation cases. For the translation vectors, we used the L2 norms of the differences. For the rotation angles, we took the angular difference, $1 - \cos(\theta - \theta^*)$, after some angular normalization steps to compensate for geometric symmetries of ellipses; we will discuss this particular issues in Section 4. Table 1 summarizes the outcomes. For higher r_X (or r_Y),

(r_X, r_Y)	$\boldsymbol{c}^*_{T_X}$	$\theta^*_{T_X}$	$\boldsymbol{c}^*_{T_Y}$	$\theta^*_{T_Y}$	$\boldsymbol{c}^*_X$	θ^*_X	$\boldsymbol{c}^*_Y$	θ^*_Y
(2.2, 2.2)	0.1221	0.0006	0.1330	0.0009	0.1756	0.0001	0.1754	0.0001
(2.2, 1.4)	0.1195	0.0011	0.1250	0.0050	1.4679	0.0151	1.2203	0.0288
(2.2, 1.1)	0.1329	0.0007	0.1208	0.0578	1.6153	0.0243	1.3037	0.0366
(1.4, 1.4)	0.1277	0.0067	0.1258	0.0056	0.2466	0.0084	0.3417	0.0007
(1.4, 1.1)	0.1196	0.0532	0.1296	0.0584	0.9755	0.0293	0.8263	0.0395
(1.1, 1.1)	0.1144	0.0393	0.1212	0.0586	0.3743	0.0096	0.4564	0.0120

Table 1: Accuracy of parameter estimation for T_X, T_Y, ϕ_X and ϕ_Y. The numbers in the table are averaged over 50 cases.

the estimation accuracy for T_X (or T_Y) increases. Note that with a higher r_X implies a clearer directionality of a primary object. The simulation outcomes explains that a clearer directionality of primary objects would help to align them and estimate T_X accurately. When r_X is below 1.4, the estimation accuracy degrades significantly. We do not suggest to apply the proposed approach for analyzing the aggregations of geometries with r_X less than 1.4. On the other hand, the estimation accuracy of ϕ_X or ϕ_Y did not depend much on r_X or r_Y.

In addition, we performed replicated experiments to see how the estimation accuracy varies over different random samples. We repeated the generative procedure (Steps 1 to 4) with fixed $r_X = 1.4$ and $r_Y = 1.1$ for 50 times to draw 50 simulation datasets, each of which contains 50 aggregation cases. For each dataset, we applied our proposed algorithm and evaluated the estimation accuracy. We computed the standard deviation of the accuracy over 50 datasets, which were 0.0123 for $\boldsymbol{c}^*_{T,X}$, 0.0002 for $\theta^*_{T,X}$, 0.0093 for $\boldsymbol{c}^*_{T,Y}$, 0.0015 for $\theta^*_{T,Y}$, 0.3425 for $\boldsymbol{c}^*_X$, 0.0117 for θ^*_X, 0.3399 for $\boldsymbol{c}^*_Y$, and 0.0118 for θ^*_Y. The variations were very small.

4 STATISTICAL ANALYSIS OF AGGREGATION

The major scientific questions that we want to answer were (1) whether there are preferential orientations of primary objects when they aggregate, and (2) if so, what the orientations are. In this section, we present a statistical analysis to answer those questions.

Suppose that we have N aggregation observations,

$$\{(X_n, Y_n, Z_n); n = 1, ...N\},$$

where X_n and Y_n are the simply connected subsets of $\mathbb{X}$ that represents two primary geometric objects for the nth observation, and Z_n is the simply connected subset of $\mathbb{X}$ that represents the corresponding aggregate. As described in Section 3.1, the $2N$ primary objects are grouped into K shape categories based on their geometric similarities, and for each shape category, we identified a reference shape and had all primary objects in the category aligned to the reference shape to define the standard coordinate systems for the primary objects.

Some shape categories may have geometrical symmetries around their major axes and minor axes, e.g., a rod and an ellipse. The major axis of a geometric object X_n is defined by the first principal loading vector of the coordinates in X_n, and the minor axis is the unit vector perpendicular to the major axis. Note that with the alignment described in Section 3.1, the major axis of a primary object is along the x-axis, and the minor axis is along the y-axis. For the primary objects belonging to a shape category symmetric around the major and minor axis, the following orientation angles of the primary objects are indistinguishable due to the geometrical symmetry,

$$\theta \equiv -\theta \equiv \pi - \theta \equiv -\pi + \theta \text{ for } \theta \in [0, \pi/2]. \tag{9}$$

Therefore, for a symmetric shape category, we normalize orientation θ to

$$\tilde{\theta} = \begin{cases} |\theta| & \text{if } |\theta| \leq \pi/2, \\ \pi - |\theta| & \text{otherwise,} \end{cases} \tag{10}$$

which is basically one of the θ's equivalent forms in the first quadrant $[0, \pi/2]$.

We work with the observations of θ for a non-symmetric category or the observations of $\tilde{\theta}$ for a symmetric shape category for necessary statistical inferences. The probability distribution of θ for a non-symmetric case can be modeled as a von Mises distribution, which is popularly used to describe a unimodal probability density of angular data (Mardia et al. 2012). The statistical inferences on the distribution model have been well studied in

circular statistics (Fisher 1995); therefore, we will not reiterate them in this paper. This section focuses on statistical analysis of $\tilde{\theta}$ for symmetric cases.

For a symmetric shape category, the equivalence (9) holds in θ, and the probability density function of θ should have the following symmetries,

$$f(\theta) = f(-\theta) = f(-\pi + \theta) = f(\pi - \theta). \tag{11}$$

Therefore, if f has a mode at $\mu \in [0, \pi/2]$, it also has the modes at $-\mu$, $-\pi + \mu$ and $\pi - \mu$. A von-Mises distribution is popularly used to describe a unimodal probability density of angular data (Mardia et al. 2012). We takes a mixture of the four von Mises distributions with equal weights to represent the four modes caused by the four-way symmetry,

$$\begin{aligned}
f(\theta; \mu, \kappa) =& \frac{1}{8\pi I_0(\kappa)} \exp\{\kappa \cos(\theta - \mu)\} + \frac{1}{8\pi I_0(\kappa)} \exp\{\kappa \cos(\theta + \pi - \mu)\} \\
&+ \frac{1}{8\pi I_0(\kappa)} \exp\{\kappa \cos(\theta + \mu)\} + \frac{1}{8\pi I_0(\kappa)} \exp\{\kappa \cos(\theta - \pi + \mu)\} \\
=& \frac{1}{8\pi I_0(\kappa)} \exp\{\kappa \cos(\theta - \mu)\} + \frac{1}{8\pi I_0(\kappa)} \exp\{-\kappa \cos(\theta - \mu)\} \\
&+ \frac{1}{8\pi I_0(\kappa)} \exp\{\kappa \cos(\theta + \mu)\} + \frac{1}{8\pi I_0(\kappa)} \exp\{-\kappa \cos(\theta + \mu)\} \\
=& \frac{1}{4\pi I_0(\kappa)} \cosh(\kappa \cos(\theta - \mu)) + \frac{1}{4\pi I_0(\kappa)} \cosh(\kappa \cos(\theta + \mu)) \\
=& \frac{1}{2\pi I_0(\kappa)} \cosh(\kappa \cos(\mu) \cos(\theta)) \cosh(\kappa \sin(\mu) \sin(\theta)),
\end{aligned}$$

where $\cosh(\cdot)$ is a hyperbolic cosine function, and $\mu \in [0, \pi/2]$. One can easily check that the density function satisfies the symmetry (11) as desired. Note that the normalization (10) applies for mirroring θ onto the first quadrant $[0, \pi/2]$, and f has the same density for all quadrants. Therefore, the density function of the normalized angle $\tilde{\theta}$ is simply the four times of f,

$$g(\tilde{\theta}; \mu, \kappa) = \frac{2}{\pi I_0(\kappa)} \cosh(\kappa \cos(\mu) \cos(\tilde{\theta})) \cosh(\kappa \sin(\mu) \sin(\tilde{\theta})), \tag{12}$$

where $\mu, \tilde{\theta} \in [0, 2\pi]$. One can show $\int_0^{\pi/2} g(\tilde{\theta}; \mu, \kappa) = 1$, so it is a valid probability density function. The two parameters μ and κ can be estimated by the maximum likelihood estimation described in Section 4.1, and the goodness-of-fit test for the estimated parameters can be performed by the method described in Section 4.2. Sections 4.3 and 4.4 describes the statistical hypotheses testing problems to test the two scientific hypotheses that we mentioned in the beginning of this section.

4.1 Maximum Likelihood Estimation

We present a numerical procedure for the maximum likelihood estimates of μ and κ for $g(\tilde{\theta};\mu,\kappa)$ given a random sample $\{\tilde{\theta}_1,\ldots,\tilde{\theta}_N\}$ from the density. The log likelihood function is

$$L_N(\mu,\kappa)=\sum_{n=1}^{N}\log(\cosh(\kappa\cos(\mu)\cos(\tilde{\theta}_n)))+\log(\cosh(\kappa\sin(\mu)\sin(\tilde{\theta}_n)))-N\log(I_0(\kappa)).$$

The first order necessary condition, $\frac{\partial L_N}{\partial\mu}=0$ and $\frac{\partial L_N}{\partial\kappa}=0$, does not give a closed form expression for μ and κ. The two parameters μ and κ can be numerically optimized by the Newton-Raphson algorithm using the first order derivatives and the second order derivatives of the log likelihood function. The expressions for the first and second order derivatives can be found in the online supplementary material, and the initial solution for μ can be specified to the sample angular mean μ_0, and the initial solution κ_0 can be found using the unbiased estimator of $\frac{I_1(\kappa)}{I_0(\kappa)}$,

$$\mu_0=\arctan\left(\frac{\bar{s}}{\bar{c}}\right)\ \text{and}\ \frac{I_1(\kappa_0)}{I_0(\kappa_0)}=\frac{N}{N-1}\bar{c}^2+\bar{s}^2-\frac{1}{N-1},$$

where $\bar{s}=\frac{1}{N}\sum_{n=1}^{N}\sin(\tilde{\theta}_n)$ and $\bar{c}=\frac{1}{N}\sum_{n=1}^{N}\cos(\tilde{\theta}_n)$.

The maximum likelihood estimation procedure was tested for three simulation cases. We first drew a random sample of size 1000 from $g(\tilde{\theta};\mu,\kappa)$ with μ and κ specified in Table 2 and used the random sample to estimate μ and κ as described in this section. The estimates $\hat{\mu}$ and $\hat{\kappa}$ were compared to the values of μ and κ used as simulation inputs, and the differences were evaluated. The differences were averaged over 100 replicated experiments, which provided the biases of the estimates, and the variance of the estimates were averaged over the replicated runs. Table 2 summarizes the outcomes.

4.2 Goodness-of-Fit Test

We use the Kolmogorov-Smirnov test (Arnold & Emerson 2011) to test the goodness-of-fit of $g(\tilde{\theta};\hat{\mu},\hat{\kappa})$ to a random sample $\{\tilde{\theta}_1,\ldots,\tilde{\theta}_N\}$. Let $G(\tilde{\theta})$ denote the cumulative distribution function for $g(\tilde{\theta};\hat{\mu},\hat{\kappa})$ and $G_n(\tilde{\theta})$ denote the empirical cumulative distribution function,

$$G_n(\tilde{\theta})=\frac{1}{N}\sum_{n=1}^{N}I_{[-\infty,\tilde{\theta}]}(\tilde{\theta}_n).$$

Simulation Inputs	$\mu = \pi/6, \kappa = 10$		$\mu = \pi/4, \kappa = 10$		$\mu = \pi/6, \kappa = 5$	
	$\hat{\mu}$	$\hat{\kappa}$	$\hat{\mu}$	$\hat{\kappa}$	$\hat{\mu}$	$\hat{\kappa}$
Bias	0.00240	0.46050	0.00012	0.0192	0.0107	0.4912
Variance	0.00008	0.28300	0.00009	0.2353	0.00023	0.1653

Table 2: Biases and variances of the maximum likelihood estimates $\hat{\mu}$ and $\hat{\kappa}$. Each value in the table is the average value over 100 replicated runs.

The test statistic is

$$T_N = \sqrt{n} \sup_{\tilde{\theta}} |G(\tilde{\theta}) - G_n(\tilde{\theta})|.$$

If the test statistic is below a critical value $t_{\alpha,N}$, the fit of G to G_n is good. The critical value can be achieved by the Monte Carlo simulation,

Step 1. Take a random sample of size N from $g(\tilde{\theta}; \hat{\mu}, \hat{\kappa})$, and get the empirical cumulative distribution function G_n for the random sample.

Step 2. Compute T_N.

Step 3. Repeat Step 1 and Step 2 many times and get $1 - \alpha$ quantile of the resulting T_N values, which becomes $t_{\alpha,N}$.

4.3 Testing the Uniformity of Distribution

The first hypothesis to test is whether there is a preferential orientation of a primary object in its aggregate. The hypothesis can be tested using the following statistical hypothesis test on $g(\tilde{\theta}; \mu, \kappa)$,

$$\text{H0:} g(\tilde{\theta}; \mu, \kappa) \text{ is uniform.}$$

$$\text{H1:} g(\tilde{\theta}; \mu, \kappa) \text{ is not uniform.}$$

Note that as κ decreases, the density function $g(\tilde{\theta}; \mu, \kappa)$ becomes closer to an angular uniform distribution and becomes perfectly uniform with $\kappa = 0$ and nearly uniform with $\kappa \leq 0.5$. We formulate the uniformity testing as testing on κ

$$\text{H0:} \kappa \leq 0.5$$

$$\text{H1:} \kappa > 0.5.$$

We use the likelihood ratio test to test the hypothesis. The likelihood ratio test statistic for testing H0 versus H1,

$$R_\kappa = \max_{\kappa>0.5} L_N(\mu, \kappa) - \max_{\kappa\leq 0.5} L_N(\mu, \kappa).$$

The two maximization problems in R_κ can be evaluated by maximizing the likelihood function with linear constraints on κ. The critical value of the test statistic can be easily determined using the Monte Carlo simulation similar to the one described in Section 4.2.

4.4 Testing the Mean Orientation

The second hypothesis to test is whether the mean orientation of a primary object in its aggregate is μ_0. This test can be formulated as testing on the mean parameter μ,

$$\text{H0:}\mu = \mu_0$$

$$\text{H1:}\mu \neq \mu_0.$$

It can be tested using the likelihood ratio test with the difference of two log likelihoods as a test statistic,

$$R_\mu = \max_{\mu,\kappa} L_N(\mu, \kappa) - \max_{\mu=\mu_0,\kappa} L_N(\mu, \kappa).$$

The two maximization problems in R_μ can be evaluated by maximizing the likelihood function with linear constraints on μ. The critical value of the test statistic can be easily determined using the Monte Carlo simulation similar to the one described in Section 4.2.

5 APPLICATION TO NANOPARTICLE AGGREGATION

The motivating example described in Section 2 provided 184 aggregation observations for nanoparticles, i.e., $N = 184$. Section 3.1 was applied to group the $2N$ primary objects into K shape categories by their geometric similarities; $K = 3$ was chosen by the AIC. For each shape category, we identified a reference shape and had the primary objects in the category aligned to the reference shape. Figure 5 illustrates the images of the primary particles after the alignment. Notably, the major axes of the primary particles were aligned

to the horizontal line (i.e. x-axis), which indicates that the alignment task worked well. Apparently those three shape categories are distinct in terms of an aspect ratio, which is defined as the ratio of the major axis length and the minor axis length of a shape. The mean aspect ratios are 1.99 for the first category, 1.40 for the second, and 1.22 for the last category. Based on the typical appearances of nanoparticles, we named shape category 1 as 'Rod' ($k = 1$, 82 objects), shape category 2 as 'Ellipse' ($k = 2$, 146 objects), and shape category 3 as 'NearSphere' ($k = 3$, 140 objects).

The N aggregation observations can be classified into six groups, depending on the shape categories of the primary objects involved in the aggregations, Rod-Rod (12 cases), Rod-Ellipse (26 cases), Rod-NearSphere (32 cases), Ellipse-Ellipse (33 cases), Ellipse-NearSphere (54 cases), and NearSphere-NearSphere (27 cases). We achieved the orientation angles of primary nanoparticles normalized to $[0, \pi/2]$ as described in Section 4,

$$\{(\tilde{\theta}_X^{(n)}, \tilde{\theta}_Y^{(n)}); n = 1, ...N\},$$

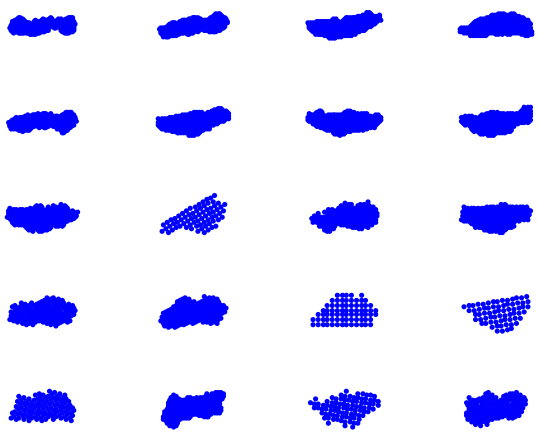
where $(\tilde{\theta}_X^{(n)}, \tilde{\theta}_Y^{(n)})$ are the orientation angles of two primary particles for the nth observation.

We first looked at the angular correlation coefficients of $\tilde{\theta}_X^{(n)}$ and $\tilde{\theta}_Y^{(n)}$ for each aggregation group. Let $N_{k1,k2}$ denote the collection of observation indices n's that correspond to aggregations of shape categories $k1$ and $k2$. Following Fisher & Lee (1983), the angular correlation coefficient $\rho_{k1,k2}$ is defined as,

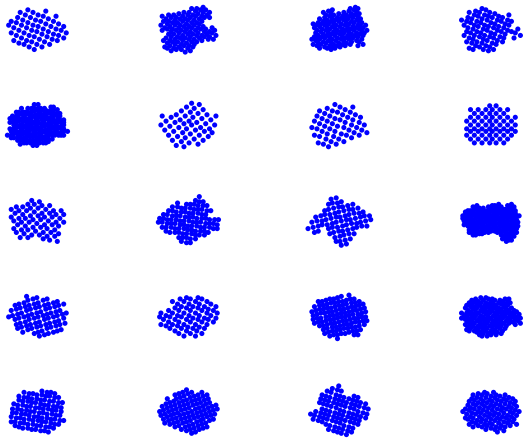
$$\rho_{k1,k2} = \frac{\sum_{i,j \in N_{k1,k2}} \sin(\tilde{\theta}_X^{(i)} - \tilde{\theta}_X^{(j)}) \sin(\tilde{\theta}_Y^{(i)} - \tilde{\theta}_Y^{(j)})}{\sqrt{\sum_{i,j \in N_{k1,k2}} \sin^2(\tilde{\theta}_X^{(i)} - \tilde{\theta}_X^{(j)})} \sqrt{\sum_{i,j \in N_{k1,k2}} \sin^2(\tilde{\theta}_Y^{(i)} - \tilde{\theta}_Y^{(j)})}}.$$

The corresponding coefficient of determination, $\rho^2_{k1,k2}$, is 0.1859 for Rod-Rod, 0.0273 for Rod-Ellipse, 0.0937 for Rod-NearSphere, 0.1195 for Ellipse-Ellipse, 0.0008 for Ellipse-NearSphere, 0.2252 for NearSphere-NearSphere. When $k1 = k2$, the coefficients were computed in between $\min\{\tilde{\theta}_X^{(n)}, \tilde{\theta}_Y^{(n)}\}$ and $\max\{\tilde{\theta}_X^{(n)}, \tilde{\theta}_Y^{(n)}\}$. Typically, $\rho^2_{k1,k2}$ less than 0.3 is regarded nearly uncorrelated, so $\tilde{\theta}_X^{(n)}$ and $\tilde{\theta}_Y^{(n)}$ are nearly linearly independent for the six aggregation groups.

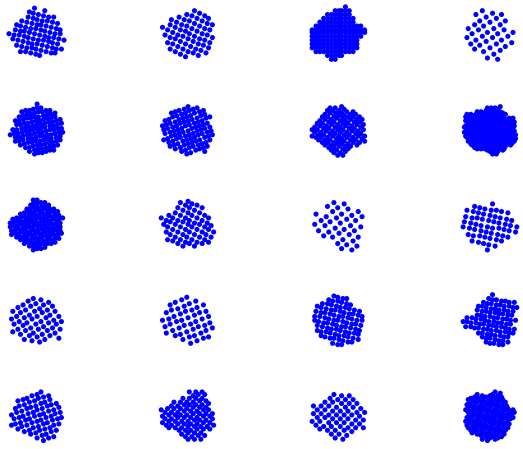
Given the nearly independence of $\tilde{\theta}_X^{(n)}$ and $\tilde{\theta}_Y^{(n)}$ and a limited number of observations per group, we approximately model the joint distribution of the two angles with a product of the marginal distributions of the two angles. Let $p_{k1|k2}(\tilde{\theta})$ denote the marginal density



(a) Shape Category 1: Rod



(b) Shape Category 2: Ellipse



(c) Shape Category 3: NearSphere

Figure 5: Alignment outcomes for three shape categories

function of $\tilde{\theta}$ of shape category $k1$ when it aggregates with shape category $k2$, which is assume to be

$$p_{k1|k2}(\tilde{\theta}) = g(\tilde{\theta}; \mu_{k1,k2}, \kappa_{k1,k2}).$$

The maximum likelihood estimation procedure described in Section 4.1 was applied for $k1 = 1, 2$ and $k2 = 1, 2, 3$. We have not analyzed $k1 = 3$ cases (Near-Sphere cases) because the cases are subject to significant estimation errors as we showed from the simulation study in Section 3.3. Let $\hat{\mu}_{k1,k2}$ and $\hat{\kappa}_{k1,k2}$ denote the estimated $\mu_{k1,k2}$ and $\kappa_{k1,k2}$. Figure 6 shows the $p_{k1|k2}(\tilde{\theta})$ with $\hat{\mu}_{k1,k2}$ and $\hat{\kappa}_{k1,k2}$. Section 4.2 was applied for the goodness-of-fit testing of

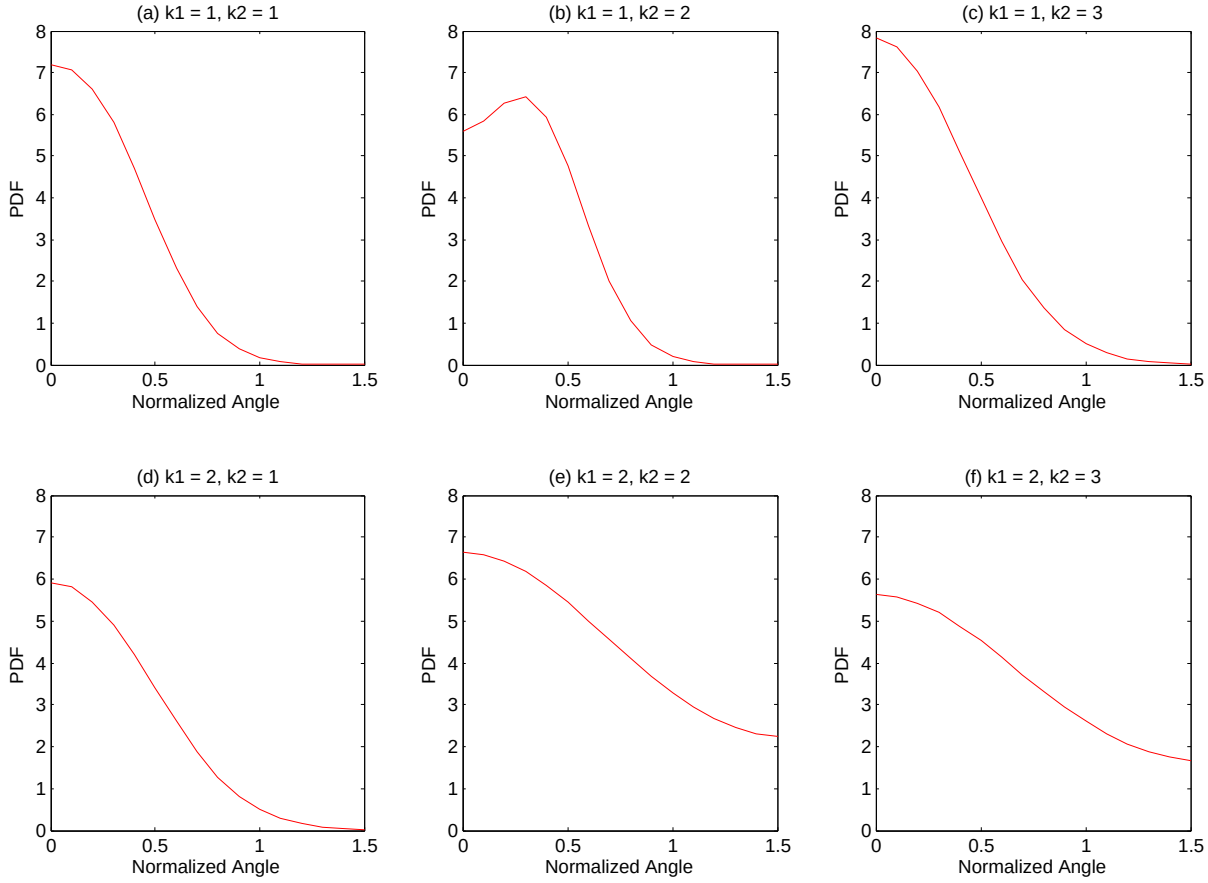


Figure 6: Estimated probability density functions (PDFs)

the estimated density functions. For all cases, the estimated CDFs were very comparable to the corresponding empirical CDFs, and the goodness-of-fit test also showed no significant difference between them with 95% significance level. Figure 7 shows the cumulative density

functions (G) of the estimated PDFs, with comparisons to the empirical CDFs (G_n).

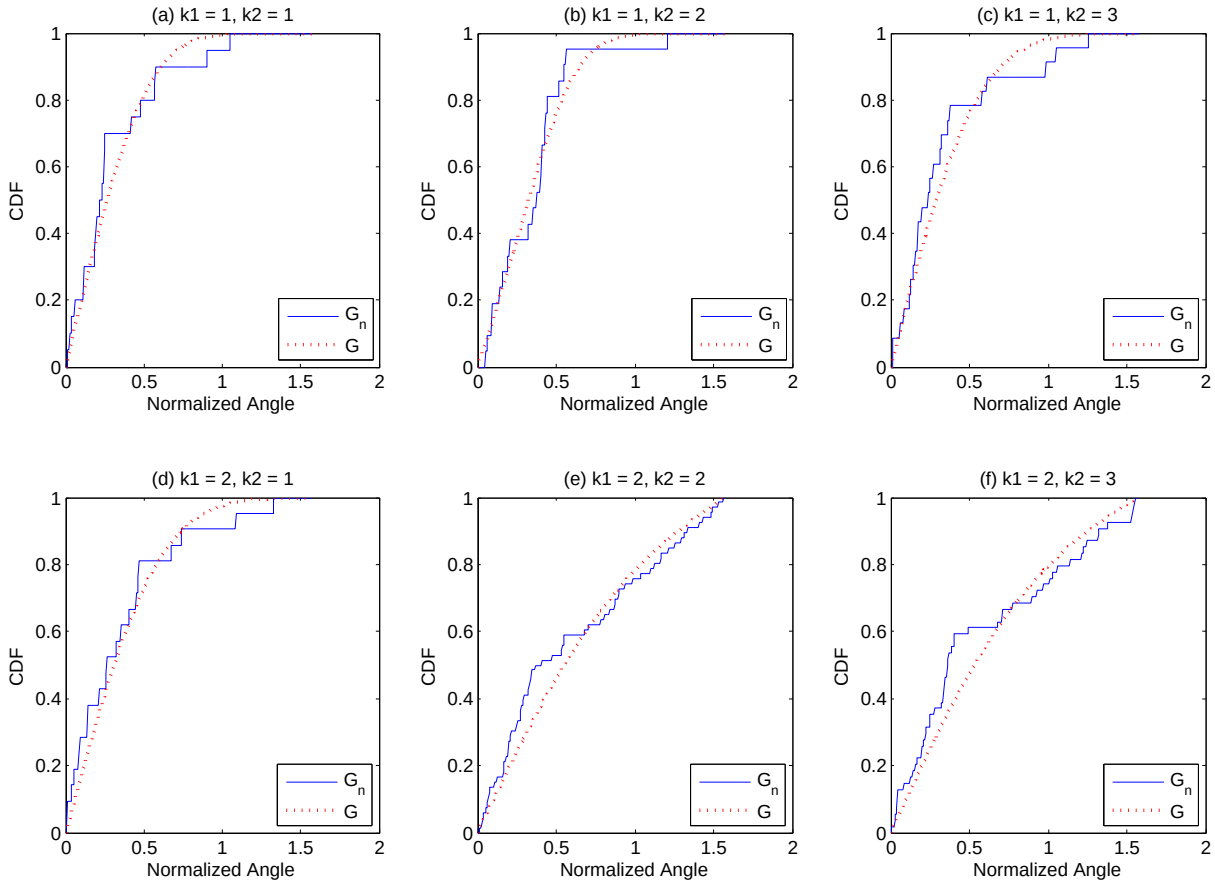


Figure 7: Goodness-of-Fit Test; G denote the estimated CDF, and G_n denotes the empirical CDF.

We also tested a scientific hypothesis related to whether there is a preferential orientation of shape category $k1$ when it aggregates with shape category $k2$. We applied Section 4.3 to test

$$\text{H0:} p_{k1|k2}(\theta) \text{ is uniform.}$$

$$\text{H1:} p_{k1|k2}(\theta) \text{ is not uniform.}$$

With 95% significance level, the null hypothesis was rejected for $(k1, k2) = (1, 1)$, $(1, 2)$, $(1, 3)$ $(2, 1)$, $(2, 2)$ and $(2, 3)$. The results indicate strong evidences that rod-like and ellipse-like nanoparticles have preferential orientations when they aggregate with rod-like, ellipse-like or near-sphere like nanoparticles.

We performed a steered molecular dynamics (SMD) simulation of a rod-to-rod particle aggregation (Welch et al. 2016), which allowed us to compute the energy barriers against aggregation for different orientations of rods. According to the simulation, when the major axes of two aggregating rods were not oriented toward the aggregation center, the compression of solvent monolayers at rod surfaces significantly increased when the rods became close to each other. The increase of the solvation force placed a large energy barrier against the aggregation of the two rods. The energy barrier was minimized when both of the rods' major axes were oriented toward the aggregation center. This implies the preferential orientation of a rod particle in its aggregate is zero. Note that the direction of the major axis is zero. To test how our experimental observations are consistent with the simulation result, we formulated a hypothesis testing problem, which basically examines whether the mean orientation $\mu_{1,1}$ for a Rod-Rod aggregation is zero,

$$\text{H0:}\mu_{1,1} = 0$$

$$\text{H1:}\mu_{1,1} \neq 0.$$

We applied Section 4.4 to test the hypothesis. With 95% significant level, the null hypothesis cannot be rejected. In other words, with high significance, the experimental observations are consistent with the output of the SMD simulation.

6 CONCLUSION

We have presented a statistical model for studying the oriented attachment of nanoparticles with dynamic microscopy data, i.e., studying the preferential orientations of two primary nanoparticles participating in the particle aggregation. We geometrically defined a particle aggregation by two primary geometries merging into a secondary geometry. Each primary geometry in dynamic microscopy data was represented by a simply connected subset in a two-dimensional Euclidean space with a certain choice of its standard coordinate system, and the secondary geometry was represented by a union of the two primary geometries having certain orientations. We proposed a shape alignment approach to define the orientations of the primary geometries within the secondary geometry, and presented a numerical algorithm for solving the approach. We believe that the work for mathematically formu-

lating and analyzing particle aggregations has not been previously performed. We also presented a statistical model to describe the probability distribution of the orientations of primary geometries in their aggregates and formulated several statistical hypothesis testing problems.

We applied our proposed method to our motivating example of nanoparticle aggregations. The results demonstrated that two primary particles were aligned along certain preferential orientations during their aggregation and the orientations were consistent with what we achieved from a molecular dynamics simulation. By far, the microscopy and nanoscience community has been manually cherry-picking and analyzing individual cases of nanoparticle aggregation. To the best of our knowledge, our study is the first attempt to statistically analyze multiple cases of nanoparticle aggregations from a single nanoparticle synthesis process.

SUPPLEMENTARY MATERIAL

Implementation details: a pdf file containing some implementation details of the proposed method, including Section A. the optimization algorithm to solve problems (4) and (7), and Section B. the first and second order derivatives of the log likelihood function in Section 4.1.

References

Akaike, H. (1992), Information theory and an extension of the maximum likelihood principle, *in* 'Breakthroughs in statistics', Springer, pp. 610–624.

Arnold, T. B. & Emerson, J. W. (2011), 'Nonparametric goodness-of-fit tests for discrete null distributions', *The R Journal* **3**(2), 34–39.

Dryden, I. & Mardia, K. (1998), *Statistical shape analysis*, Wiley.

Fisher, N. I. (1995), *Statistical analysis of circular data*, Cambridge University Press.

Fisher, N. I. & Lee, A. (1983), 'A correlation coefficient for circular data', *Biometrika* pp. 327–332.

Green, P. J. & Mardia, K. V. (2006), 'Bayesian alignment using hierarchical models, with applications in protein bioinformatics', *Biometrika* **93**(2), 235–254.

Kendall, D. G. (1984), 'Shape manifolds, procrustean metrics, and complex projective spaces', *Bulletin of the London Mathematical Society* **16**(2), 81–121.

Lele, S. (1993), 'Euclidean distance matrix analysis (edma): estimation of mean form and mean form difference', *Mathematical Geology* **25**(5), 573–602.

Li, D., Nielsen, M. H., Lee, J. R., Frandsen, C., Banfield, J. F. & De Yoreo, J. J. (2012), 'Direction-specific interactions control crystal growth by oriented attachment', *Science* **336**(6084), 1014–1018.

Mardia, K. V., Kent, J. T., Zhang, Z., Taylor, C. C. & Hamelryck, T. (2012), 'Mixtures of concentrated multivariate sine distributions with applications to bioinformatics', *Journal of Applied Statistics* **39**(11), 2475–2492.

Mémoli, F. (2007), On the use of Gromov-Hausdorff distances for shape comparison, *in* 'Eurographics Symposium on Point-based Graphics', The Eurographics Association, pp. 81–90.

Mémoli, F. & Sapiro, G. (2005), 'A theoretical and computational framework for isometry invariant recognition of point cloud data', *Foundations of Computational Mathematics* **5**(3), 313–347.

Mora, C. & Kwan, A. (2000), 'Sphericity, shape factor, and convexity measurement of coarse aggregate for concrete using digital image processing', *Cement and concrete research* **30**(3), 351–358.

Park, C., Woehl, T. J., Evans, J. E. & Browning, N. D. (2015), 'Minimum cost multi-way data association for optimizing large-scale multitarget tracking of interacting objects', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**(3), 611–624.

Rangarajan, A., Chui, H. & Bookstein, F. L. (1997), The softassign procrustes matching algorithm, *in* 'Information Processing in Medical Imaging', Springer, pp. 29–42.

Schmidler, S. C. (2007), 'Fast bayesian shape matching using geometric algorithms', *Bayesian statistics* **8**, 471–490.

Srivastava, A., Klassen, E., Joshi, S. H. & Jermyn, I. H. (2011), 'Shape analysis of elastic curves in Euclidean spaces', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(7), 1415–1428.

Wang, W. (1999), 'Image analysis of aggregates', *Computers & Geosciences* **25**(1), 71–81.

Welch, D. A., Woehl, T., Park, C., Faller, R., Evans, J. E. & Browning, N. D. (2016), 'Understanding the role of solvation forces on the preferential attachment of nanoparticles in liquid', *ACS Nano* **10**(1), 181–187.

Woehl, T., Evans, J., Arslan, I., Ristenpart, W. & Browning, N. (2012), 'Direct *in situ* determination of the mechanisms controlling nanoparticle nucleation and growth', *ACS Nano* **6**(10), 8599–8610.

Younes, L. (1998), 'Computable elastic distances between shapes', *SIAM Journal on Applied Mathematics* **58**(2), 565–586.

Zhang, W., Crittenden, J., Li, K. & Chen, Y. (2012), 'Attachment efficiency of nanoparticle aggregation in aqueous dispersions: modeling and experimental validation', *Environmental Science & Technology* **46**(13), 7054–7062.