

γ -SUP: A CLUSTERING ALGORITHM FOR CRYO-ELECTRON MICROSCOPY IMAGES OF ASYMMETRIC PARTICLES

BY TING-LI CHEN*, DAI-NI HSIEH*, HUNG HUNG[†], I-PING TU*, PEI-SHIEN WU[‡], YI-MING WU*, WEI-HAU CHANG* AND SU-YUN HUANG*

*Academia Sinica**, *National Taiwan University[†]* and *Duke University[‡]*

Cryo-electron microscopy (cryo-EM) has recently emerged as a powerful tool for obtaining three-dimensional (3D) structures of biological macromolecules in native states. A minimum cryo-EM image data set for deriving a meaningful reconstruction is comprised of thousands of randomly orientated projections of identical particles photographed with a small number of electrons. The computation of 3D structure from 2D projections requires clustering, which aims to enhance the signal to noise ratio in each view by grouping similarly oriented images. Nevertheless, the prevailing clustering techniques are often compromised by three characteristics of cryo-EM data: high noise content, high dimensionality and large number of clusters. Moreover, since clustering requires registering images of similar orientation into the same pixel coordinates by 2D alignment, it is desired that the clustering algorithm can label misaligned images as outliers. Herein, we introduce a clustering algorithm γ -SUP to model the data with a q -Gaussian mixture and adopt the minimum γ -divergence for estimation, and then use a self-updating procedure to obtain the numerical solution. We apply γ -SUP to the cryo-EM images of two benchmark macromolecules, RNA polymerase II and ribosome. In the former case, simulated images were chosen to decouple clustering from alignment to demonstrate γ -SUP is more robust to misalignment outliers than the existing clustering methods used in the cryo-EM community. In the latter case, the clustering of real cryo-EM data by our γ -SUP method eliminates noise in many views to reveal true structure features of ribosome at the projection level.

1. Introduction and motivating example. Determining 3D atomic structure of large biological molecules is important for elucidating the physicochemical mechanisms underlying vital processes. In 2006, the Nobel Prize in chemistry was awarded to structural biologist Roger D. Kornberg for his studies of the molecular basis of eukaryotic transcription, in which he obtains for the first time an actual picture at the molecular level of the structure of RNA polymerase II during the stage of actively making messenger RNA. Three years later in 2009, the same prize went to three X-ray crystallographers, Venkatraman Ramakrishnan, Thomas A. Steitz and Ada E. Yonath, for their revelation at the atomic level of the structure and workings

Received September 2012; revised August 2013.

Key words and phrases. Clustering algorithm, cryo-EM images, γ -divergence, k -means, mean-shift algorithm, multilinear principal component analysis, q -Gaussian distribution, robust statistics, self-updating process.

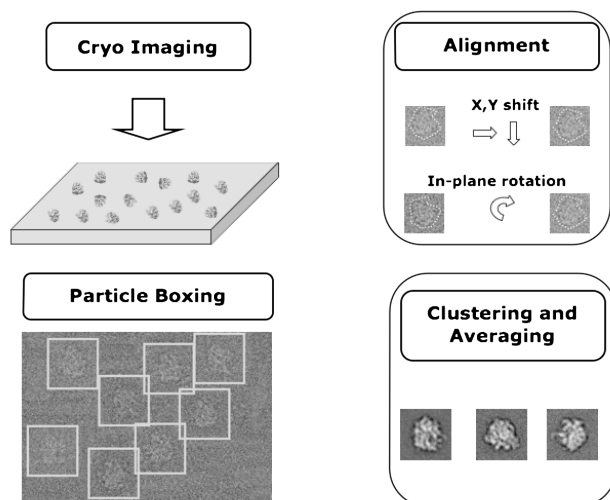


FIG. 1. *The flowchart for cryo-EM analysis.*

of the ribosome, with its even larger and more complex machinery that translates the information contained in RNA into a poly-peptide chain. Despite these successes, most large proteins have resisted all attempts at crystallization. This has led to the emergence of cryo-electron microscopy (cryo-EM), an alternative to X-ray crystallography for obtaining 3D structures of macromolecules, since it can focus electrons to form images without the need of crystals [Henderson (1995), Saibil (2000), van Heel et al. (2000), Frank (2002, 2009, 2012, Grassucci, Taylor and Frank (2011), Jiang et al. (2008), Liu et al. (2010)].

A flowchart for cryo-EM analysis is shown in Figure 1. In the sample preparation step, macromolecules obtained through biochemical purification in the native condition are frozen so rapidly that the surrounding water forms a thin layer of amorphous ice to embed the molecules [Adrian et al. (1984), Dubochet (2012), Lepault, Booy and Dubochet (1983)]. The data obtained by cryo-EM imaging consists of a large number of particle images representing projections from random orientations of the macromolecule. An essential step in reconstructing the 3D structure from these images is to determine the 3D angular relationship of the 2D projections, which is a challenging task for the following reasons. First, the images are of poor quality, as they are heavily contaminated by shot noise induced by the extremely low number of electrons used to prevent radiation damage. Second, in contrast to X-ray crystallography, which restricts the conformation of the macromolecule in the crystal, cryo-EM data retain any conformation variations of the macromolecule that exist in their original solution state, which means the data is a mixture of conformations on top of the viewing angles. The left panel of Figure 1 explains how the 2D cryo-EM images are collected and the right panel shows the commonly used strategy to improve the signal-to-noise ratio (SNR) of those images.

Given reasonably clear 2D projections, an ab initio approach based on the projection-slice theorem for 3D reconstruction from 2D projections is available [Bracewell (1956)]. This theorem states that any two nonparallel 2D projections of the same 3D object would share a common line in Fourier space. This common-line principle underlies the first 3D reconstruction of a spherical virus with icosahedral symmetry from electron microscope micrographs [Crowther et al. (1970)]. The same principle was further applied to the problem of angular reconstruction of the asymmetric particle [van Heel (1987)]. In practice, a satisfactory solution depends on the quality of the images and becomes increasingly unreachable for raw cryo-EM images as the SNR gets too low. It is therefore necessary to enhance the SNR of each view by averaging many well aligned cryo-EM images coming from a similar viewing angle. Clustering is thus aimed at grouping together the cryo-EM images with nearly the same viewing angle. This step requires pre-aligning the images because incorrect in-plane rotations and shiftings would prevent successful clustering. Failure to cluster images into homogeneous groups would, in turn, render the determination of the 3D structure unsatisfactory. Currently, most approaches for clustering cryo-EM images are rooted in the k -means method, which has been found to be unsatisfactory [Yang et al. (2012)].

Here, we focus on the clustering step and assume that the image alignment has been carried through. In the vast number of clustering algorithms developed, two major approaches are taken. A model-based approach [Banfield and Raftery (1993)] models the data as a mixture of parametric distributions and the mean estimates are taken to be the cluster centers. A distance-based approach enforces some “distance” to measure the similarity between data points, with notable examples being hierarchical clustering [Hartigan (1975)], the k -means algorithm [Lloyd (1982), McQueen (1967)] and the SUP clustering algorithm [Chen and Shiu (2007), Shiu and Chen (2012)]. In this paper, we combine these two approaches to propose a clustering algorithm γ -SUP. We model the data with a q -Gaussian mixture [Amari and Ohara (2011), Eguchi, Komori and Kato (2011)] and adopt the γ -divergence [Cichocki and Amari (2010), Eguchi, Komori and Kato (2011), Fujisawa and Eguchi (2008)] for measuring the similarity between the empirical distribution and the model distribution, and then borrow the self-updating procedure from SUP [Chen and Shiu (2007), Shiu and Chen (2012)] to obtain a numerical solution. While minimizing the γ -divergence leads to a soft rejection in the sense that the estimate downweights the deviant points, the q -Gaussian mixture helps set a rejection region when the deviation gets too large. Both of these factors resist outliers and contribute robustness to our clustering algorithm. To execute the self-updating procedure, we start with treating each individual data point as the cluster representative of a singleton cluster and, in each iteration, we update the cluster representatives through the derived estimating equations until all the representatives converge. This self-updating procedure ensures that neither knowledge of the number of clusters nor random initial centers are required.

To investigate how γ -SUP would perform when applied to cryo-EM images, we tested two sets of cryo-EM images. The first set, consisting of noisy simulated RNA polymerase II images of different views projected from a defined orientation, was chosen in order to decouple the alignment issues from clustering issues, allowing for quantitative comparison between γ -SUP and other clustering methods. The second set consisted of 5000 real cryo-EM images of ribosome bound with an elongation factor that locks it into a defined conformation. For the test on the simulated data, both perfectly aligned cases and misaligned cases were examined. γ -SUP did well in separating different views in which the images were perfectly aligned and was able to identify most of the deliberately misaligned images as outliers. For the ribosome images, γ -SUP was successful in that the cluster averages were consistent with the views projected from the known ribosome structure.

The paper is organized as follows. Section 2 reviews the concepts of γ -divergence and the q -Gaussian distribution, which are the core components of γ -SUP. In Section 3 we develop our γ -SUP clustering algorithm from the perspective of the minimum γ -divergence estimation of the q -Gaussian mixture model. The performance of γ -SUP is further evaluated through simulations in Section 4, and through a set of real cryo-EM images in Section 5. The paper ends with a conclusion in Section 6.

2. A review of γ -divergence and the q -Gaussian distribution. In this section we briefly review the concepts of γ -divergence and the q -Gaussian distribution, which are the key technical tools for our γ -SUP clustering algorithm.

2.1. *γ -divergence.* The most widely used divergence of distributions is probably the Kullback-Leibler divergence (KL-divergence) due to its connection to maximum likelihood estimation (MLE). The γ -divergence, indexed by a power parameter $\gamma > 0$, is a generalization of KL-divergence. Let

$$\mathcal{M} = \left\{ f : 0 < \int_{\mathcal{X}} f^{\gamma+1} < \infty, f \geq 0 \right\},$$

where $f : \mathcal{X} \subset \mathbb{R}^n \mapsto \mathbb{R}^+$ is a nonnegative function defined on \mathcal{X} . For simplicity, we assume \mathcal{X} is either a discrete set or a connected region.

DEFINITION 1 [Cichocki and Amari (2010), Eguchi, Komori and Kato (2011), Fujisawa and Eguchi (2008)]. For $f, g \in \mathcal{M}$, define the γ -divergence $D_\gamma(\cdot \| \cdot)$ and γ -cross entropy $C_\gamma(\cdot \| \cdot)$ as follows:

$$(1) \quad \begin{aligned} D_\gamma(f \| g) &= C_\gamma(f \| g) - C_\gamma(f \| f) \\ \text{with } C_\gamma(f \| g) &= -\frac{1}{\gamma(\gamma + 1)} \int \frac{g^\gamma(x)}{\|g\|_{\gamma+1}^\gamma} f(x) dx, \end{aligned}$$

where $\|g\|_{\gamma+1} = \{ \int g^{\gamma+1}(x) dx \}^{1/(\gamma+1)}$ is a normalizing constant.

The γ -divergence can be understood as the divergence function associated with a specific scoring function, namely, the pseudospherical score [Gneiting and Raftery (2007), Good (1971)]. The pseudospherical score is given by $S(f, x) = f^\gamma(x)/\|f\|_{\gamma+1}^\gamma$. The associated divergence function between f and g can be calculated from equation (7) in Gneiting and Raftery (2007) to be

$$d(f\|g) = \int S(f, x)f(x)dx - \int S(g, x)f(x)dx = \gamma(\gamma + 1)D_\gamma(f\|g).$$

This implies that $d(\cdot\|\cdot)$ and $D_\gamma(\cdot\|\cdot)$ are equivalent. Moreover, $D_\gamma(\cdot\|\cdot)$ can also be expressed as a functional Bregman divergence [Frigyik, Srivastava and Gupta (2008)] by taking $\Phi(f) = \|f\|_{\gamma+1}$. The corresponding Bregman divergence is

$$\begin{aligned} D_\Phi(f\|g) &= \Phi(f) - \Phi(g) - \delta\Phi[g, f - g] = \|f\|_{\gamma+1} - \int \frac{g^\gamma(x)}{\|g\|_{\gamma+1}^\gamma} f(x)dx \\ &= \gamma(\gamma + 1)D_\gamma(f\|g), \end{aligned}$$

where $\delta\Phi[g, h]$ is the Gâteaux derivative of Φ at g along direction h . Note that $\|g\|_{\gamma+1}$ is a normalizing constant so that the cross entropy enjoys the property of being projective invariant, that is, $C_\gamma(f\|cg) = C_\gamma(f\|g)$, $\forall c > 0$ [Eguchi, Komori and Kato (2011)]. By Hölder's inequality, it can be shown that, for $f, g \in \Omega$ (defined below), $D_\gamma(f\|g) \geq 0$ and equality holds if and only if $g = \lambda f$ for some $\lambda > 0$ [Eguchi, Komori and Kato (2011)]. Thus, by fixing a scale, for example,

$$\Omega = \{f \in \mathcal{M} : \|f\|_{\gamma+1} = 1\},$$

D_γ defines a legitimate divergence on Ω . There are other possible ways of fixing a scale, for example, $\Omega = \{f \in \mathcal{M} : \int f(x)dx = 1\}$.

In the limiting case, $\lim_{\gamma \rightarrow 0} D_\gamma(f\|g) = D_0(f\|g) = \int f(x) \ln\{f(x)/g(x)\} dx$, which gives the KL-divergence. The MLE, which corresponds to the minimization of the KL-divergence $D_0(\cdot\|\cdot)$, has been shown to be optimal in parameter estimation in many settings in the sense of having minimum asymptotic variance. This optimality comes with the cost that the MLE relies on the correctness of model specification. Therefore, the MLE or the minimization of the KL-divergence may not be robust against model deviation and outlying data. On the other hand, the minimum γ -divergence estimation is shown to be robust [Fujisawa and Eguchi (2008)] against data contamination. It is this robustness property that makes γ -divergence suitable for the estimation of mixture components, where each component is a local model [Mollah et al. (2010)].

2.2. The q -Gaussian distribution. The q -Gaussian distribution is a generalization of the Gaussian distribution obtained by replacing the usual exponential function with the q -exponential

$$\exp_q(u) = \{1 + (1 - q)u\}_+^{1/(1-q)} \quad \text{where } \{x\}_+ = \max\{x, 0\}.$$

In this article, we adopt $q < 1$, which corresponds to q -Gaussian distributions with compact support. We will explain this necessary condition of compact support later.

Let \mathcal{S}_p denote the collection of all strictly positive definite $p \times p$ symmetric matrices.

DEFINITION 2 [Modified from Amari and Ohara (2011), Eguchi, Komori and Kato (2011)]. For a fixed $q < 1 + \frac{2}{p}$, define the p -variate q -Gaussian distribution $G_q(\mu, \Sigma)$ with parameters $\theta = (\mu, \Sigma) \in \mathbb{R}^p \times \mathcal{S}_p$ to have the probability density function (p.d.f.),

$$(2) \quad f_q(x; \theta) = \frac{c_{p,q}}{(\sqrt{2\pi})^p \sqrt{|\Sigma|}} \exp_q \{u(x; \theta)\}, \quad x \in \mathbb{R}^p,$$

where $u(x; \theta) = -\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)$ and $c_{p,q}$ is a constant so that $\int f_q(x; \theta) dx = 1$.

The constant $c_{p,q}$ is given below [cf. Eguchi, Komori and Kato (2011)]:

$$(3) \quad c_{p,q} = \begin{cases} \frac{(1-q)^{p/2} \Gamma(1 + p/2 + 1/(1-q))}{\Gamma(1 + 1/(1-q))}, & \text{for } -\infty < q < 1, \\ 1, & \text{for } q \rightarrow 1, \\ \frac{(q-1)^{p/2} \Gamma(1/(q-1))}{\Gamma(1/(q-1) - p/2)}, & \text{for } 1 < q < 1 + \frac{2}{p}. \end{cases}$$

The class of the q -Gaussian distributions covers some well-known distributions. In the limit as q approaches 1, the q -Gaussian distribution reduces to the Gaussian distribution. For $1 < q < 1 + \frac{2}{p}$, the q -Gaussian distribution becomes the multivariate t -distribution. This can be seen by setting $v = 2/(q - 1) - p > 0$. Then, $f_q(x; \theta)$ in (2) is proportional to

$$(4) \quad \left\{ 1 + \frac{1}{v}(x - \mu)^T \left(\frac{p+v}{v} \Sigma \right)^{-1} (x - \mu) \right\}^{-(p+v)/2},$$

which is exactly the p.d.f. of a p -variate t -distribution (up to a constant term) with location and scale parameters $(\mu, \frac{p+v}{v} \Sigma)$ and degrees of freedom v . Depending on the choice of q , the support of $G_q(\mu, \Sigma)$ also differs. For $1 + \frac{2}{p} > q \geq 1$ (i.e., for the Gaussian distribution and t -distribution), the support of $G_q(\mu, \Sigma)$ is the entire \mathbb{R}^p . For $q < 1$, however, the support of $G_q(\mu, \Sigma)$ is compact and depends on q in the form

$$(5) \quad \left\{ x : (x - \mu)^T \Sigma^{-1}(x - \mu) < \frac{2}{1-q} \right\}.$$

Thus, choosing $q < 1$ leads to zero mutual influence between clusters in our clustering algorithm. Note that if $X \sim G_q(\mu, \Sigma)$ with $q < 1 + \frac{2}{p+2}$,¹ then $E(X) = \mu$ and $\text{Cov}(X) = \frac{2}{2+(p+2)(1-q)}\Sigma$.

2.3. *Minimum γ -divergence for estimating a q -Gaussian.* The γ -divergence is a discrepancy measure for two functions in \mathcal{M} . Its minimum can then be used as a criterion to approximate an underlying p.d.f. f from a certain model class \mathcal{M}_Θ parameterized by $\theta \in \Theta \subset \mathbb{R}^m$. It has been deduced that minimizing $D_\gamma(f \| g)$ over g is equivalent to minimizing the γ -loss function [Fujisawa and Eguchi (2008)]

$$(6) \quad L_{\gamma, f}(g) = -\frac{1}{\gamma} \ln \left\{ \int g^\gamma(x) f(x) dx \right\} + \frac{1}{\gamma + 1} \ln \left\{ \int g^{\gamma+1}(x) dx \right\}.$$

Thus, at the population level, f is estimated by

$$(7) \quad f^* = \underset{g \in \mathcal{M}_\Theta}{\text{argmin}} D_\gamma(f \| g) = \underset{g \in \mathcal{M}_\Theta}{\text{argmin}} L_{\gamma, f}(g).$$

At this moment, we consider \mathcal{M}_Θ to be the family of q -Gaussian distributions $G_q(\mu, \Sigma)$ with $\theta = (\mu, \Sigma)$. Then, for any given values of γ and q , the loss function $L_{\gamma, f}(g)$ evaluated at $g(x) = f_q(x; \theta) \in \mathcal{M}_\Theta$ becomes

$$\begin{aligned} L_{\gamma, f}\{f_q(x; \theta)\} &= -\frac{1}{\gamma} \ln \left[\int f(x) \left\{ \frac{\exp_q(u(x; \theta))^{\gamma+1}}{\int \{\exp_q(u(v; \theta))\}^{\gamma+1} dv} \right\}^{\gamma/(\gamma+1)} dx \right] \\ &= -\frac{1}{\gamma} \ln \left[\int f(x) \left\{ f_{(\gamma+q)/(\gamma+1)} \left(x; \mu, \frac{1}{\gamma+1} \Sigma \right) \right\}^{\gamma/(\gamma+1)} dx \right]. \end{aligned}$$

Direct calculation then gives that minimizing $L_{\gamma, f}\{f_q(x; \theta)\}$ over possible values of θ is equivalent to maximizing

$$(8) \quad \int f(x) |\Sigma|^{-(1/2)(\gamma/(\gamma+1))} [\exp_q\{u(x; \theta)\}]^\gamma dx.$$

For high-dimensional data, however, it is impractical to estimate the covariance matrix Σ and its inverse. Since our main interest is to find cluster centers, we employ $\Sigma = \sigma^2 I_p$ as our working model. By taking the derivative of (8) with respect to μ , we get the stationary equation for the maximizer μ^* for any fixed σ^2 :

$$(9) \quad \begin{aligned} \mu^* &= \frac{\int x f(x) [\exp_q\{u(x; \mu^*, \sigma^2)\}]^{\gamma-(1-q)} dx}{\int f(x) [\exp_q\{u(x; \mu^*, \sigma^2)\}]^{\gamma-(1-q)} dx} \\ &= \frac{\int x w(x; \mu^*, \sigma^2) dF(x)}{\int w(x; \mu^*, \sigma^2) dF(x)}, \end{aligned}$$

¹For $q < 1 + \frac{2}{p+2}$, it ensures the existence of the v th moment of X .

where $w(x; \mu^*, \sigma^2) = [\exp_q\{u(x; \mu^*, \sigma^2)\}]^{\gamma-(1-q)}$ is the weight function and $F(x)$ is the cumulative distribution function corresponding to $f(x)$.

Given the observed data $\{x_i\}_{i=1}^n$, the sample analogue of μ^* can be obtained naturally by replacing $F(x)$ in (9) with the empirical distribution function $\hat{F}(x)$ of $\{x_i\}_{i=1}^n$. This gives the stationarity condition for μ^* at the sample level:

$$(10) \quad \mu^* = \frac{\sum_{i=1}^n x_i w(x_i; \mu^*, \sigma^2)}{\sum_{i=1}^n w(x_i; \mu^*, \sigma^2)}.$$

One can see that the weight function w assigns the contribution of x_i to μ^* . Thus, a robust estimator should encourage the property that smaller weight is given to those x_i farther away from μ^* and zero weight to extreme outliers. These can be achieved by choosing proper values of (γ, q) in w . In particular, when $q < 1$, we have from (5) that

$$(11) \quad w(x; \mu^*, \sigma^2) = \begin{cases} \left(1 - \frac{1-q}{2\sigma^2} \|x - \mu^*\|_2^2\right)^{(\gamma-(1-q))/(1-q)}, & \text{for } \|x - \mu^*\|_2^2 < \frac{2\sigma^2}{1-q}, \\ 0, & \text{for } \|x - \mu^*\|_2^2 \geq \frac{2\sigma^2}{1-q}. \end{cases}$$

That is, data points outside the range $\mu^* \pm \frac{2\sigma^2}{1-q}$ do not have any influence on μ^* . Note also that when $\gamma = 1 - q$, then $w(x; \mu^*, \sigma^2) = 1$ and, thus, μ^* in (10) becomes the sample mean $n^{-1} \sum_{i=1}^n x_i$, which is not robust to outliers. This fact suggests that we should use a γ value that is greater than $1 - q$.

3. γ -SUP. In this section we introduce our clustering method, γ -SUP, which is derived from minimizing γ -divergence under a q -Gaussian mixture model.

3.1. *Model specification and estimation.* Suppose we have collected data $\{x_i\}_{i=1}^n$ with empirical probability mass $\hat{f}(x)$ and empirical c.d.f. $\hat{F}(x) = \frac{1}{n} \sum_{j=1}^n I(x_j \leq x)$, where “ \leq ” is understood componentwise. The goal is to group them into K clusters, where K is unknown and should be determined from the data. Assume that f is a mixture of K components with p.d.f.

$$(12) \quad f(x) = \sum_{k=1}^K \pi_k f_q(x; \theta_k),$$

where each component is modeled by a q -Gaussian distribution indexed by $\theta_k = (\mu_k, \sigma^2)$. Most model-based clustering approaches (e.g., those assuming a Gaussian mixture model) aim to learn the whole model f by minimizing a divergence (e.g., KL-divergence) between f and the empirical probability mass \hat{f} . They therefore suffer the problem of having to specify the number of components K before

implementation. To overcome this difficulty, instead of minimizing $D_\gamma(\hat{f}\|f)$ to learn f directly, we consider learning each component $f_q(\cdot; \theta_k)$ of f separately through the minimization problem

$$(13) \quad \min_{\theta} D_\gamma\{\hat{f}\|f_q(\cdot; \theta)\}.$$

The validity of (13) to learning all components of f relies on the locality of γ -divergence, as shown in Lemma 3.1 of Fujisawa and Eguchi (2008). The authors have proven that, at the population level, $D_\gamma\{f\|f_q(\cdot; \theta)\}$ is approximately proportional to $D_\gamma\{f_q(\cdot; \theta_k)\|f_q(\cdot; \theta)\}$, provided that the model $f_q(x; \theta)$ and the remaining components $\{f_q(x; \theta_\ell) : \ell \neq k\}$ are well separated. We also refer the readers to Mollah et al. (2010) for a comprehensive discussion about the locality of γ -divergence. Consequently, we are motivated to find all local minimizers of (13), each of which corresponds to an estimate of one component of f . Moreover, the number of local minimizers provides an estimate of K . A detailed implementation algorithm that finds all local minimizers and estimates K is introduced in the next subsection.

3.2. Implementation: Algorithm and tuning parameters. We have shown in Section 2.3 that, for any given σ^2 , solving (13) is equivalent to finding the cluster center μ^* that satisfies the stationary equation (10). Starting with a set of initial cluster centers $\{\hat{\mu}_i^{(0)}\}$ indexed by i , we consider the following fixed-point algorithm to solve (10):

$$(14) \quad \hat{\mu}_i^{(\ell+1)} = \frac{\int x w(x; \hat{\mu}_i^{(\ell)}, \sigma^2) d\hat{F}(x)}{\int w(x; \hat{\mu}_i^{(\ell)}, \sigma^2) d\hat{F}(x)}, \quad \ell = 0, 1, 2, \dots$$

Multiple initial centers are necessary to find multiple solutions of (10). To avoid the problem of random initial centers, in this paper we consider the natural choice

$$(15) \quad \{\hat{\mu}_i^{(0)} = x_i\}_{i=1}^n.$$

Other choices are possible, but (15) gives a straightforward updating path $\{\hat{\mu}_i^{(\ell)} : \ell = 0, 1, 2, \dots\}$ for each observation x_i . At convergence, the distinct values of $\{\hat{\mu}_i^{(\infty)}\}_{i=1}^n$ provide estimates of the cluster centers $\{\mu_k\}_{k=1}^K$ and the number of clusters. Moreover, cases whose updating paths converge to the same cluster center are clustered together. Though derived from a minimum γ -divergence perspective, we note that (14) combined with (15) has the same form as the mean-shift clustering [Fukunaga and Hostetler (1975)] for mode seeking. It turns out that clustering through (14)–(15) shares the same properties as mean-shift clustering. Our γ -SUP framework provides the mean-shift algorithm an information theoretic justification, namely, it minimizes the γ -divergence under a q -Gaussian mixture model. We call (14)–(15) the “nonblurring γ -estimator,” to distinguish it from our main proposal introduced in the next paragraph.

While the nonblurring mean-shift updates the cluster centers with the original data points being fixed [which corresponds to a fixed \hat{F} in (14) over iterations], the blurring mean-shift [Cheng (1995)] is a variant of the nonblurring mean-shift algorithm, which updates the cluster centers and moves (i.e., blurs) the data points simultaneously. Shiu and Chen (2012) proposed self-updating process (SUP) as a clustering algorithm. The blurring mean-shift can be viewed as an SUP with a homogeneous updating rule (SUP allows nonhomogeneous updating). It has been reported [Shiu and Chen (2012)] that SUP possesses many advantages, especially in the presence of outliers. Thus, we are motivated to implement the minimum γ -divergence estimation via an SUP-like algorithm and call it γ -SUP. In particular, the γ -SUP algorithm is constructed by replacing $\hat{F}(x)$ in (14) with $\hat{F}^{(\ell)}(x) = \frac{1}{n} \sum_{j=1}^n I(\hat{\mu}_j^{(\ell)} \leq x)$, to reflect its nature in updating blurred data as model representatives over iterations. That is, after one iteration, we believe that the blurred data is more representative of the population than the previous one and, thus, the blurred data is treated as the new sample to be entered into the next iteration. This gives the self-updating process of γ -SUP:

$$(16) \quad \hat{\mu}_i^{(\ell+1)} = \frac{\int x w(x; \hat{\mu}_i^{(\ell)}, \sigma^2) d\hat{F}^{(\ell)}(x)}{\int w(x; \hat{\mu}_i^{(\ell)}, \sigma^2) d\hat{F}^{(\ell)}(x)}, \quad \ell = 0, 1, 2, \dots$$

The update (16) can be expressed as, for $i = 1, \dots, n$,

$$(17) \quad \hat{\mu}_i^{(1)} = \frac{\sum_{j=1}^n w_{ij}^{(0)} \hat{\mu}_j^{(0)}}{\sum_{j=1}^n w_{ij}^{(0)}} \rightarrow \hat{\mu}_i^{(2)} = \frac{\sum_{j=1}^n w_{ij}^{(1)} \hat{\mu}_j^{(1)}}{\sum_{j=1}^n w_{ij}^{(1)}} \rightarrow \dots \rightarrow \hat{\mu}_i^{(\infty)},$$

where

$$(18) \quad \begin{aligned} w_{ij}^{(\ell)} &= \left\{ \exp_q \left(-\frac{1}{2\sigma^2} \|\hat{\mu}_i^{(\ell)} - \hat{\mu}_j^{(\ell)}\|_2^2 \right) \right\}^{\gamma - (1-q)} \\ &= \exp_{1-s} \left(-\frac{1}{\tau^2} \|\hat{\mu}_i^{(\ell)} - \hat{\mu}_j^{(\ell)}\|_2^2 \right) \end{aligned}$$

with the scale parameter τ and the model parameter s being

$$(19) \quad \tau = \sigma \sqrt{\frac{2}{\gamma - (1-q)}} \quad \text{and} \quad s = \frac{1-q}{\gamma - (1-q)} > 0.$$

Here $s > 0$ is a consequence of choosing $q < 1$ and $\gamma > 1 - q$ as mentioned in the end of Section 2.3. Thus, γ -SUP involves only (s, τ) as the tuning parameters. It has been found in our numerical studies that γ -SUP is quite insensitive to the choice of s and that τ plays the decisive role in the performance of γ -SUP. We thus suggest a choice of a small positive value of s , say, 0.025, in practical implementation. A phase transition plot (Figure 5) will be introduced to determine τ in Section 4.

TABLE 1
 γ -SUP clustering algorithm

Inputs: Data matrix $X \in \mathbb{R}^{n \times p}$, n instances with p variables;
Tuning parameters (s, τ) .
Outputs: Number of clusters K and cluster centers $\{\hat{\mu}_k\}_{k=1}^K$;
Cluster membership assignment $\{c_i\}_{i=1}^n$ for each of $\{x_i\}_{i=1}^n$.

begin
 lter $\leftarrow 0$
 start with: $\tilde{\mu}_i \leftarrow x_i/\tau, i = 1, \dots, n$
 repeat
 for $i = 1 : n$
 $w_{ij} \leftarrow \exp_{1-s}(-\|\tilde{\mu}_i - \tilde{\mu}_j\|_2^2), j = 1, \dots, n$
 $z_i \leftarrow \sum_{j=1}^n \frac{w_{ij}}{\sum_{k=1}^n w_{ik}} \tilde{\mu}_j$
 end
 $\tilde{\mu}_i \leftarrow z_i, i = 1, \dots, n$
 lter \leftarrow lter + 1
 until convergence
 output distinct cluster centers $\{\tau \tilde{\mu}_i, 1 \leq i \leq n\}$ and cluster membership
end

Note: The parameter τ is linearly proportional to the influence region radius that defines the similarity inside a cluster.

It can be seen from (17) that, in updating $\hat{\mu}_i^{(\ell)}$ in the ℓ th iteration, γ -SUP takes a weighted average over the candidate model representatives $\{\hat{\mu}_j^{(\ell)}\}_{j=1}^n$ according to the weights $\{w_{ij}^{(\ell)}\}_{j=1}^n$. Due to the weights in (18) being nonnegative and decreasing with respect to the distance $\|\hat{\mu}_i^{(\ell)} - \hat{\mu}_j^{(\ell)}\|_2$ and the compact support of the q -Gaussian distribution, the convergence of γ -SUP is assured [Chen (2013)]. We can express the weights (18) as

$$(20) \quad w_{ij}^{(\ell)} = \exp_{1-s}(-\|\tilde{\mu}_i^{(\ell)} - \tilde{\mu}_j^{(\ell)}\|_2^2) \quad \text{with } \tilde{\mu}_i^{(\ell)} = \hat{\mu}_i^{(\ell)}/\tau.$$

As a result, γ -SUP starts with n (scaled) cluster centers $\{\tilde{\mu}_i^{(0)} = x_i/\tau\}_{i=1}^n$, which avoids the problem of random initial centers. Eventually, γ -SUP converges to certain K clusters, where K depends on the tuning parameters (s, τ) , but otherwise is data-driven. Moreover, we have the cluster representatives $\{\hat{\mu}_i^{(\infty)}\}_{i=1}^n$, which contain K distinct points denoted by $\{\hat{\mu}_k = \tau \tilde{\mu}_k\}_{k=1}^K$. The corresponding cluster membership for each data point is denoted by $\{c_i\}_{i=1}^n$. The detailed algorithm of γ -SUP (15)–(16) is summarized in Table 1. Note that in our proposal, we ignore the estimation of σ^2 . The main reason is that σ^2 is absorbed into the scale parameter τ defined in (19), and a phase transition plot can be used to select τ directly.

A toy example to illustrate how data points move by γ -SUP is presented in Figure 2. Two clusters with 10 data points each are sampled from the standard

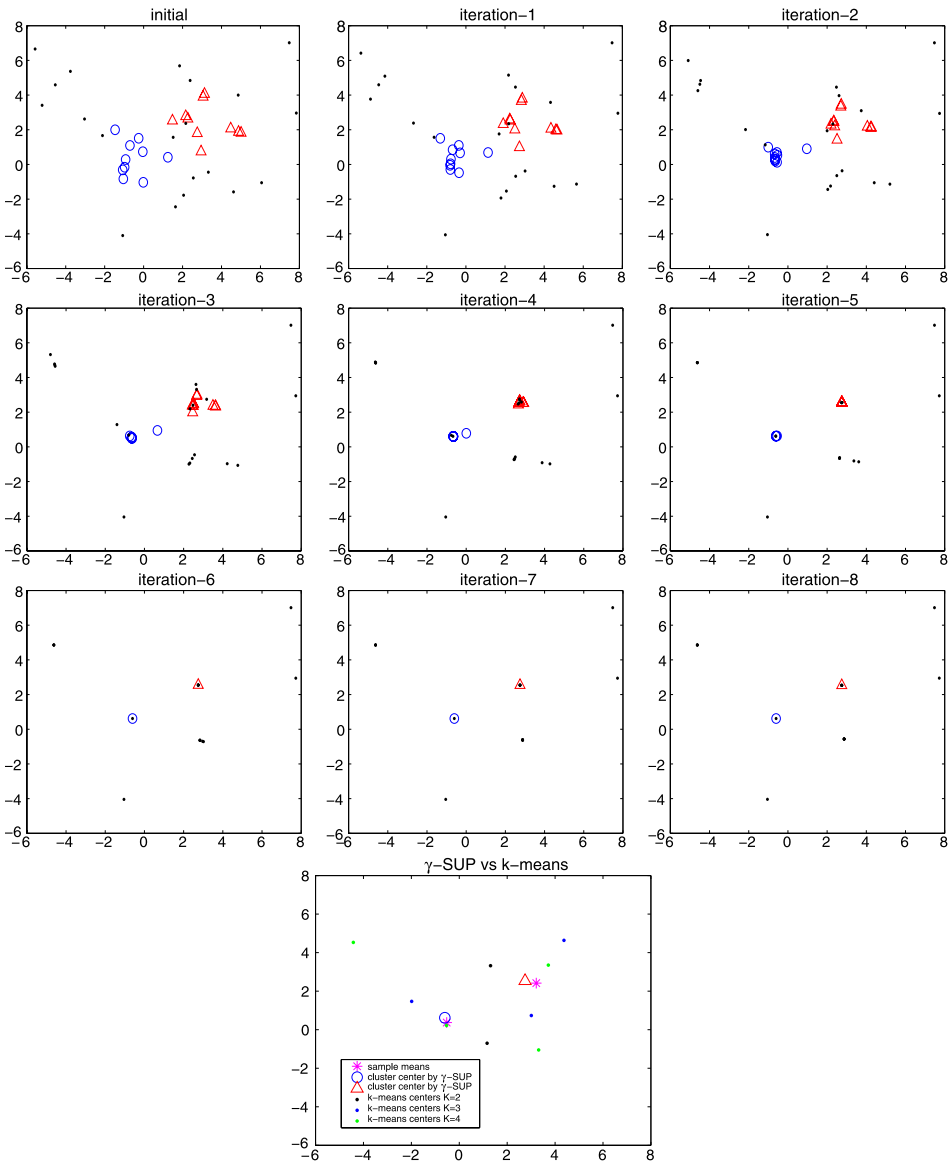


FIG. 2. How data points move by γ -SUP. The process stops at the 8th iteration. Bottom row: comparison with the sample means of the true clusters and k-means centers under $K = 2, 3, 4$.

normal distributions centered at $(0, 0)$ and $(2.355, 2.355)$, respectively. Another 20 isolated noise points are added surrounding these two clusters. The first plot (upper left) shows the initial positions of these 40 points. Then each data point is updated (blurred) according to the weighted average of its neighboring points. After the 8th iteration, no (blurred) data points move any more and the algorithm stops.

Data points with the same final position are assigned to the same cluster. There are seven clusters at the end of the self-updating process. Two cluster centers are close to the true means of the normal mixture. The other five cluster centers are formed by noise data. Data points sampled from the normal mixture are correctly merged into their target clusters. Some noise points close to these two clusters move into them, while other noise points are merged into five clusters. At the bottom row of Figure 2, we have also provided a plot for comparison with sample means and k -means centers. The sample means were computed with the information of the true cluster labels and the k -means centers were computed under given $K = 2, 3, 4$, respectively. One can see that, in the presence of outliers, cluster centers from γ -SUP can still be close to the sample means, while this is not the case for k -means under every K .

3.3. *Characteristics of γ -SUP.* Similar to (17), the nonblurring γ -estimator (14)–(15) can be expressed as

$$(21) \quad \hat{\mu}_i^{(1)} = \frac{\sum_{j=1}^n w_{ij}^{*(0)} x_j}{\sum_{j=1}^n w_{ij}^{*(0)}} \rightarrow \hat{\mu}_i^{(2)} = \frac{\sum_{j=1}^n w_{ij}^{*(1)} x_j}{\sum_{j=1}^n w_{ij}^{*(1)}} \rightarrow \dots \rightarrow \hat{\mu}_i^{(\infty)},$$

where $w_{ij}^{*(\ell)} = \exp_{1-s}(-\frac{1}{\tau^2} \|x_j - \hat{\mu}_i^{(\ell)}\|_2^2)$. Comparing (17) and (21), the constructed cluster centers from γ -SUP and the nonblurring γ -estimator are both weighted averages with the weights $w_{ij}^{(\ell)}$ and $w_{ij}^{*(\ell)}$, respectively. The principle of downweighting is important for robust model fitting [Basu et al. (1998), Field and Smith (1994), Windham (1995)]. We emphasize that the downweighting by $w_{ij}^{(\ell)}$ in (17) is *with respect to models*, since each cluster center $\hat{\mu}_i^{(\ell)}$ is a weighted average of $\{\hat{\mu}_i^{(\ell)}\}_{i=1}^n$. On the other hand, downweighting by $w_{ij}^{*(\ell)}$ in (21) is *with respect to data*, since each cluster center $\hat{\mu}_i^{(\ell)}$ is a weighted average of the original data $\{x_i\}_{i=1}^n$. The same concept can also be observed in the difference between the nonblurring mean-shift and the blurring mean-shift. It has been shown that the blurring mean-shift has a faster rate of convergence [Carreira-Perpiñán (2006), Chen (2013)]. To our knowledge, however, there is very little statistical evaluation for these two downweighting schemes in the literature. As will be demonstrated in Section 4, from a statistical perspective, downweighting with respect to models is more efficient than with respect to data in estimating the mixture model, which further supports the usage of γ -SUP in practice.

4. Numerical study.

4.1. *Synthetic data.* We show by simulation that γ -SUP is more efficient in model parameter estimation than both the nonblurring γ -estimator and the k -means estimator. Random samples of size 100 are generated from a 4-component

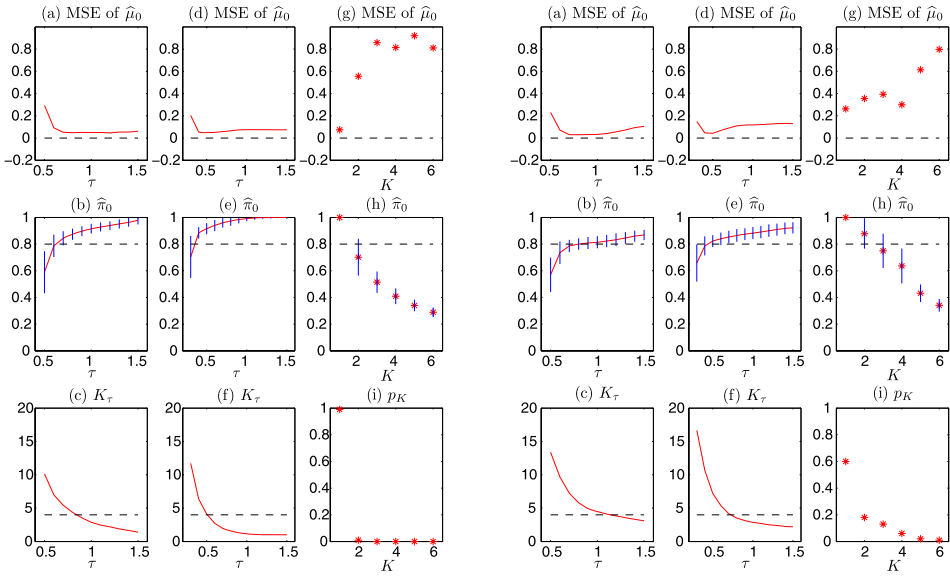


FIG. 3. Simulation results for the mixture model in (22) with $(c, \pi_0) = (2, 0.8)$ (the left panel) and $(c, \pi_0) = (4, 0.8)$ (the right panel). Results are reported for γ -SUP (the first column) and the nonblurring γ -estimator (the second column) at $s = 0.025$ and various τ values, and k -means (the third column) at various K values. (a), (d), (g): MSE of $\hat{\mu}_0$; (b), (e), (h): means of $\hat{\pi}_0$, where the vertical bars represent standard deviations; (c), (f): means of K_τ ; (i): means of p_K . The horizontal dashed line represents the target value.

normal mixture model with p.d.f.

$$(22) \quad \pi_0 f(x; \mu_0, \Sigma) + \sum_{k=1}^3 \frac{1 - \pi_0}{3} f(x; \mu_k, \Sigma),$$

where $\Sigma = I_2$, $\mu_0 = (0, 0)^T$, $\mu_1 = (c, c)^T$, $\mu_2 = (c, -2c)^T$, and $\mu_3 = (-c, 0)^T$ for some c . We set $\pi_0 = 0.8$ and treat $f(x; \mu_0, \Sigma)$ as the relevant component of interest, while $\{f(x; \mu_k, \Sigma) : k = 1, 2, 3\}$ are noise components. We apply γ -SUP and the nonblurring γ -estimator (both with $s = 0.025$) and k -means to estimate (μ_0, π_0) by using the largest cluster center and cluster size fraction as $\hat{\mu}_0$ and $\hat{\pi}_0$, respectively. The simulation results with 100 replicates under different c values are placed in Figure 3, which includes the MSE of $\hat{\mu}_0$ and the mean of $\hat{\pi}_0$ plus/minus one standard deviation. To evaluate the sensitivity of each method to the selection of τ or K , we report K_τ (the average selected number of components under τ) for γ -SUP and the nonblurring γ -estimator, and report p_K , the probability of selecting K components by the gap-statistic [Tibshirani, Walther and Hastie (2001)], for k -means.

For the case of closely spaced components, where $c = 2$, although γ -SUP and the nonblurring γ -estimator perform similarly, γ -SUP produces smaller MSE in

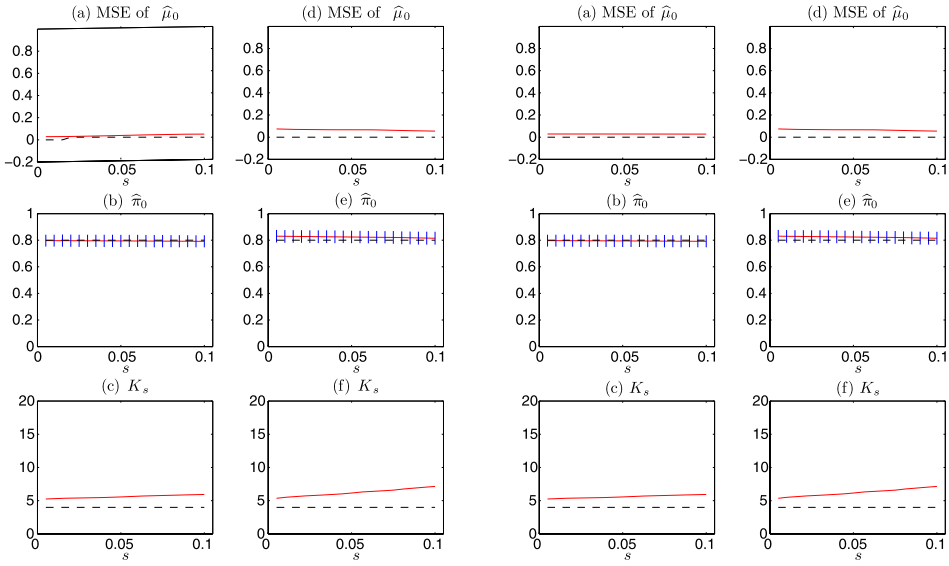


FIG. 4. Simulation results for the mixture model in (22) with $(c, \pi_0) = (2, 0.8)$ (the left panel) and $(c, \pi_0) = (4, 0.8)$ (the right panel). Results are reported for γ -SUP (the first column) and the nonblurring γ -estimator (the second column) at $\tau = 0.6$ and $s \in [0.005, 0.1]$. (a), (d): MSE of $\hat{\mu}_0$; (b), (e): means of $\hat{\pi}_0$, where the vertical bars represent standard deviations; (c), (f): mean selected number of components K_s at s . The horizontal dashed line represents the target value.

a wider range of τ values. Moreover, the mean of $\hat{\pi}_0$ from γ -SUP is closer to the target value 0.8. On the other hand, k -means provides unsatisfactory results for $K > 1$, indicating that k -means is sensitive to the selection of K when components are not well separated. The superiority of γ -SUP again has been found in the case of a moderate distance, $c = 4$, among the components. In this situation, γ -SUP still provides good estimates of (μ_0, π_0) over a wide range of $\tau \in [0.6, 1]$, while the selection of τ becomes more critical for the nonblurring γ -estimator. The minimum MSE of the nonblurring γ -estimator happens only around $\tau = 0.5$. Since τ should be determined from the data, this fact supports the applicability of γ -SUP in practical implementation. The k -means algorithm can produce satisfactory results at the true number of components $K = 4$, although its MSE is still the largest among the three methods. However, the gap-statistic does not support the choice of $K = 4$, which increases the difficulty of using a k -means with gap-statistic approach for clustering unbalanced-sized components.

We now evaluate the influence of (s, τ) to γ -SUP. The same simulation studies with $\tau = 0.6$ and $s \in [0.005, 1]$ are conducted. One can see from Figure 4 that the performance of γ -SUP and the nonblurring γ -estimator is quite insensitive to the value of s and that γ -SUP is even a bit better, which confirms that the critical tuning parameter is the scale parameter τ . As to the selection of τ , one can find an elbow-pattern (or phase transition) of K_τ in Figures 3(c) and (f). This phe-

nomenon suggests the necessity of choosing a τ value of which the number of constructed components becomes stable. It can be seen from Figure 3(a)–(b) and (d)–(e) that the suggested rule corresponds to reasonable performance for γ -SUP and the nonblurring γ -estimator, where the best result is still from γ -SUP. We will further examine this selection criterion in the next study using simulated cryo-EM images. Moreover, as discussed in Section 3.3, γ -SUP has a faster rate of convergence. Therefore, the computational time of γ -SUP is shorter than the nonblurring γ -estimator. In summary, γ -SUP performs better under different settings and is less sensitive to s -selection.

4.2. *Simulated cryo-EM images.* We use simulated cryo-EM images to demonstrate the superiority of γ -SUP in clustering homogeneous images while isolating misaligned images. A total of 128 distinct 2D images with 100×100 pixels were generated by projecting the X-ray crystal structure of RNA polymerase II filtered to 20 Angstroms in equally spaced (angle-wise) orientations.² Each image was then convoluted with the electron microscopy contrast transfer function (defocus 2 μm). Finally, 6400 images were randomly sampled with replacement from these 128 projections with i.i.d. Gaussian noise $N(0, \sigma_\varepsilon^2)$ added, where $\sigma_\varepsilon = 40, 50, 60$, so that the corresponding SNRs are 0.19, 0.12, 0.08, which reflect the low SNR nature of cryo-EM images. This procedure of simulating cryo-EM images is commonly used in the cryo-EM community [Chang et al. (2010), Hall, Nogales and Glaeser (2011), Singer et al. (2010), Sorzano et al. (2010)]. To reflect the nature of the experiment, we consider two scenarios:

1. All the images are perfectly aligned.
2. A portion of the images are misaligned.

The misaligned images are treated as outliers and should be identified individually without clustering with other correctly aligned images. Before entering the clustering algorithm, we conducted dimension reduction on this image data set. Instead of using principal component analysis (PCA), we applied multilinear principal component analysis [MPCA, Hung et al. (2012), Lu, Plataniotis and Venetsanopoulos (2008)] to reduce the dimension from 100×100 to 10×10 , as MPCA has been shown to be more efficient in dimension reduction for array data such as image sets [Hung et al. (2012)]. The extracted low-dimensional MPCA factor loadings then enter the clustering analysis.

We have tested γ -SUP, k -means, clustering 2D (CL2D) [Sorzano et al. (2010)] and k -means⁺, a variant of the latter two, for comparison. Given a prespecified number of clusters K , unlike k -means which separates the data into K clusters directly, CL2D bisects the clusters iteratively until K clusters are constructed.

²Data source: The X-ray model of RNA polymerase II is from the Protein Data Bank (PDB: 1WCM).

During the clustering process, CL2D dismisses the clusters whose size is below a certain number (30 in our case) and splits the largest cluster into two clusters once a dismiss is executed. Another difference is that CL2D adopts the correntropy [a kernel-based entropy measure, Sorzano et al. (2010)] as the measure of distance, while k -means uses the usual Euclidean norm. We thus consider k -means⁺, which is modified from CL2D by using the Euclidean norm instead. Note that CL2D, k -means⁺ and k -means require the prespecification of K and initial assignments for the cluster centers. CL2D further needs a tuning parameter for the kernel width. We have tried a few kernel widths and picked the best one for CL2D. We present the best results out of 10 runs for CL2D, k -means and k -means⁺. In contrast to this, γ -SUP is deterministic as long as its parameters s and τ are fixed. As has been demonstrated in the previous numerical studies that γ -SUP is insensitive to the s -selection, we fix $s = 0.025$. We will supply a phase transition diagram in Section 4.2.1 to choose τ .

To evaluate the performance of each method, we report the impurity and c-impurity numbers as defined below. Let $\{c_i\}$ be sets of true clusters, $\{\omega_j\}$ be sets of constructed clusters, and $|\cdot|$ be the cardinality of the set. For each output cluster ω_j , its purity number is defined by $\max_i |c_i \cap \omega_j|$. The overall purity number [Manning, Raghavan and Schtze (2008)] is the sum over all output clusters:

$$\text{purity} = \sum_j \max_i |c_i \cap \omega_j|.$$

The impurity number is defined to be

$$\text{impurity} = n - \text{purity}.$$

Note that the purity is usually defined to be the ratio of the purity number and the total number of images. Here we do not normalize it by the total number for better presentation of the simulation results. The impurity number is 0 for the perfect clustering result, but a zero impurity number does not guarantee a perfect clustering. This number cannot recognize mistakes made by splitting one class into two or more clusters. We thus define c-impurity, which is analogous to the impurity number but exchanges the roles of the true clusters and the output clusters to derive its measure of impurity. That is, define

$$\text{c-impurity} = n - \sum_i \max_j |c_i \cap \omega_j|,$$

which is able to pick up the mistakes by splitting a cluster into two or more clusters. In summary, small values of the impurity and c-impurity numbers indicate better performance of a clustering method.

TABLE 2
Clustering result with perfect alignment images

	γ -SUP	γ -SUP ⁺	CL2D	k -means ⁺	k -means
$\sigma_\varepsilon = 40$ (SNR = 0.19)					
impurity	0	0	0	0	1206
c-impurity	0	0	0	0	425
$\sigma_\varepsilon = 50$ (SNR = 0.12)					
impurity	44	0	0	34	1175
c-impurity	0	0	0	33	462
$\sigma_\varepsilon = 60$ (SNR = 0.08)					
impurity	150	0	4	0	1106
c-impurity	0	0	4	0	465

4.2.1. *Clustering with perfectly-aligned images.* Simulation results with perfectly-aligned images are placed in Table 2. For the small noise level of $\sigma_\varepsilon = 40$, γ -SUP, CL2D and k -means⁺ give perfect clustering. For larger σ_ε values, their clustering accuracies slowly decay as expected, and have comparable performances. CL2D confuses images at an impurity level of (4, 4) (which stands for impurity = 4 and c-impurity = 4) when $\sigma_\varepsilon = 60$. Quite unexpectedly, k -means⁺ confuses images at an impurity level of (34, 33) when $\sigma_\varepsilon = 50$ but performs flawlessly when $\sigma_\varepsilon = 60$. This is due to the 10 random initial centers mentioned above. The performance of k -means is poor for all σ_ε values, even when we correctly specify $K = 128$. This reflects the shortcomings of k -means when the noise level is high and when the number of clusters is large.

One can see that γ -SUP has a larger impurity number when $\sigma_\varepsilon = 60$, and we found that the errors made by γ -SUP always involve mistakenly combining two clusters as a single one. In real applications, practitioners are very likely to have prior knowledge about the expected cluster size when analyzing cryo-EM images, and it is common to further bisect an unevenly large cluster. To mimic this situation, we consider γ -SUP⁺, which modifies the result of γ -SUP by further using k -means to separate those clusters whose size is greater than 70. This threshold should be adjusted according to the ratio of the total number of images to the number of clusters expected. The results of γ -SUP⁺ are also provided in Table 2, where a perfect clustering is achieved for every σ_ε . This indicates the applicability of our proposal, as an error made by γ -SUP can be easily fixed. We remind the readers that the true number of components in this simulation is $K = 128 = 2^7$. This makes CL2D (and k -means⁺) beneficial in clustering, since the main idea of CL2D is to bisect data until a prespecified number of clusters is reached. We thus believe using γ -SUP⁺ is a fair comparison for our method. We will further see the superiority of γ -SUP in the presence of outliers (see Section 4.2.2), in which the true number of components largely exceeds 128, resulting in CL2D becoming less accurate at clustering.

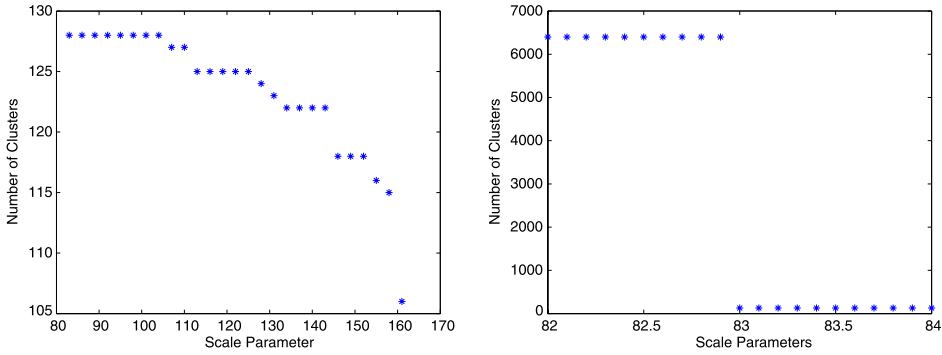


FIG. 5. The number of clusters created by γ -SUP under various values of τ . A phase transition occurs when the scale parameter τ is 83.

We demonstrate the effect of τ and provide guidance on its selection. Figure 5 gives the numbers of clusters from γ -SUP under various values of τ when $\sigma_\varepsilon = 40$, wherein we observe a phase transition in the number of clusters: γ -SUP outputs 6400 clusters when $\tau < 83$, outputs 128 clusters (a perfect result) when $\tau = 83$, and there exists no intermediate result between 128 and 6400. Moreover, the cluster number remains at 128 for quite a wide range of $\tau \in [83, 105]$. Recall that the scale parameter τ is proportional to the support region of the weight (18) and, hence, γ -SUP's updating procedure ignores the influence of data outside a certain range determined by τ . When τ is small enough that there is no influence between any two images, γ -SUP leads to 6400 clusters (i.e., each individual cryo-EM image forms one cluster). When τ reaches a critical value, the images in the same cluster can start attracting each other and will finally merge. This explains why a phase transition occurs. We observe similar phase transition phenomena for various noise structures, of which some may not happen at the perfect cluster result, but never happen far from it. Thus, the value at which the phase transition occurs can be treated as a starting value for selecting a reasonable range of τ .

4.2.2. Clustering with misaligned images. It is well possible that a set of cryo-EM images cannot be well aligned due to their low SNR. A good cluster method should be robust in the presence of misaligned images (outliers). We thus conduct simulations to evaluate the performances of clustering methods, where 10% and 20% of the images are randomly chosen to be rotated by 7.2, 14.4, 21.6, 28.8, 36 or 43.2 angular degrees ($^\circ$) clockwise. The effect of this is that each rotated image no longer shares the same signal pattern with the images in its original cluster, nor does it share a signal pattern with the other misaligned ones. An ideal outcome in this scenario would be for the clustering algorithm to treat each of these misaligned images as a singleton cluster. Including these singleton clusters, the total cluster number would become 771 for 10% misalignment and 1410 for 20%, while the meaningful cluster number would remain at 128.

TABLE 3
Clustering result with 10% misalignment images

	γ -SUP	γ -SUP ⁺	CL2D	k -means ⁺	k -means
$\sigma_\varepsilon = 40$ (SNR = 0.19)					
impurity	0	0	643	789	407
c-impurity	0	0	0	0	3452
$\sigma_\varepsilon = 50$ (SNR = 0.12)					
impurity	83	0	644	788	410
c-impurity	0	0	1	4	3482
$\sigma_\varepsilon = 60$ (SNR = 0.08)					
impurity	190	7	644	779	423
c-impurity	0	0	1	2	3541

Simulation results are presented in Table 3 for the 10% misalignment scenario and Table 4 for the 20% misalignment scenario. It can be seen that γ -SUP performs best in comparison with CL2D, k -means⁺ and k -means for all σ_ε values. Moreover, combined with the prior knowledge about expected cluster size, the result from γ -SUP can be further improved by γ -SUP⁺. These outcomes support the applicability of γ -SUP in the presence of outliers. As shown in Table 3 for the very low SNR case with $\sigma_\varepsilon = 60$, γ -SUP⁺ has only 7 misaligned images which have been wrongly assigned to true clusters, while the other misaligned images were correctly separated out as outliers. On the other hand, k -means⁺ and CL2D can do nothing about the misaligned images which are 643 ($= 771 - 128$) in the 10% misalignment case, such that their default mistakes are 643 in the impurity category. The main reason is that CL2D will still allocate outliers into certain clusters and, hence, the resulting cluster centers will still be subject to the influence of outliers. In contrast, γ -SUP allows for singleton clusters as indicated in Table 3, and the resulting cluster means are, therefore, more robust.

The performance of CL2D and k -means⁺ is more seriously impacted when the misaligned proportion is 20% (Table 4). While CL2D makes no other mistake at $\sigma_\varepsilon = 40$ and makes (1, 1) mistakes at $\sigma_\varepsilon = 50$ and $\sigma_\varepsilon = 60$ in addition to the default 643 misalignment images in the 10% case, it makes from 421 to 444 more image merges in addition to the default 1282 ($= 1410 - 128$) ones for the 20% case. The k -means⁺ method shows similar patterns in both the 10% and the 20% case. This indicates that, as a large number of outliers are forced to enter the clusters, their cluster representatives tend to be contaminated to an extent that it becomes difficult to identify their correct cluster members. In contrast, γ -SUP provides a solution to this commonly encountered situation.

5. Real cryo-EM image clustering. Ribosome is the cellular machinery that synthesizes proteins. The structure of ribosome has been intensively investigated

TABLE 4
Clustering result with 20% misalignment images

	γ -SUP	γ -SUP ⁺	CL2D	k -means ⁺	k -means
			$\sigma_\varepsilon = 40$ (SNR = 0.19)		
impurity	0	0	1703	1773	820
c-impurity	0	0	4	3	3883
			$\sigma_\varepsilon = 50$ (SNR = 0.12)		
impurity	36	1	1743	1760	833
c-impurity	0	0	1	11	3899
			$\sigma_\varepsilon = 60$ (SNR = 0.08)		
impurity	214	11	1726	1713	824
c-impurity	0	0	6	1	3909

by X-ray crystallography and cryo-EM. For the latter, the structure is obtained by 3D reconstruction using many cryo-EM images. Those images represent projections from different viewing angles. To test the performance of γ -SUP in grouping experimental data into similar views, a set of E. coli 70s ribosome was downloaded³ (60 sample images are shown in Figure 6). This data set contains 5000 single-particle cryo-EM images of randomly oriented ribosome carrying an elon-

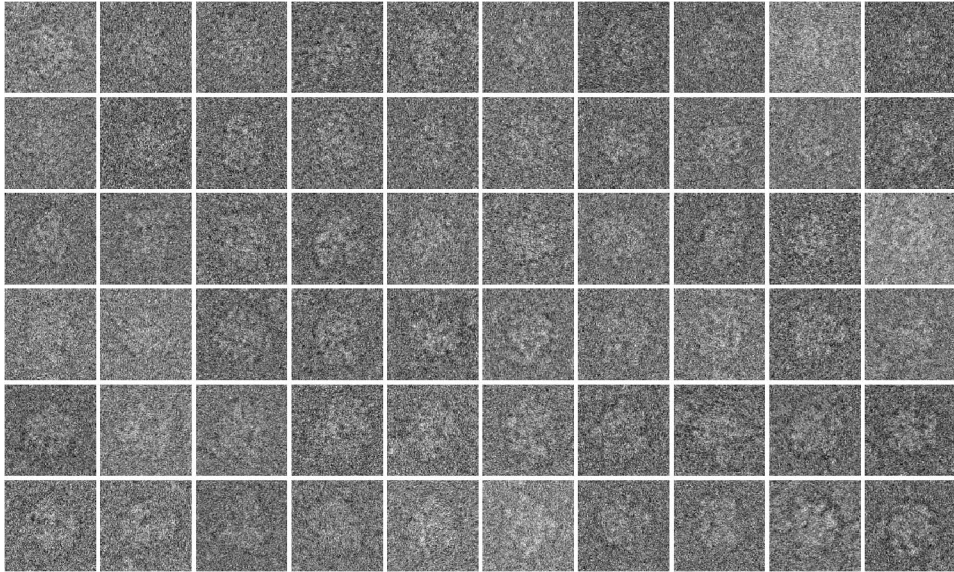


FIG. 6. 60 images are randomly chosen from the 5000 ribosome cryo-EM image set.

³The website link is http://www.ebi.ac.uk/pdbe/emdb/data/SPIDER_FRANK_data/.

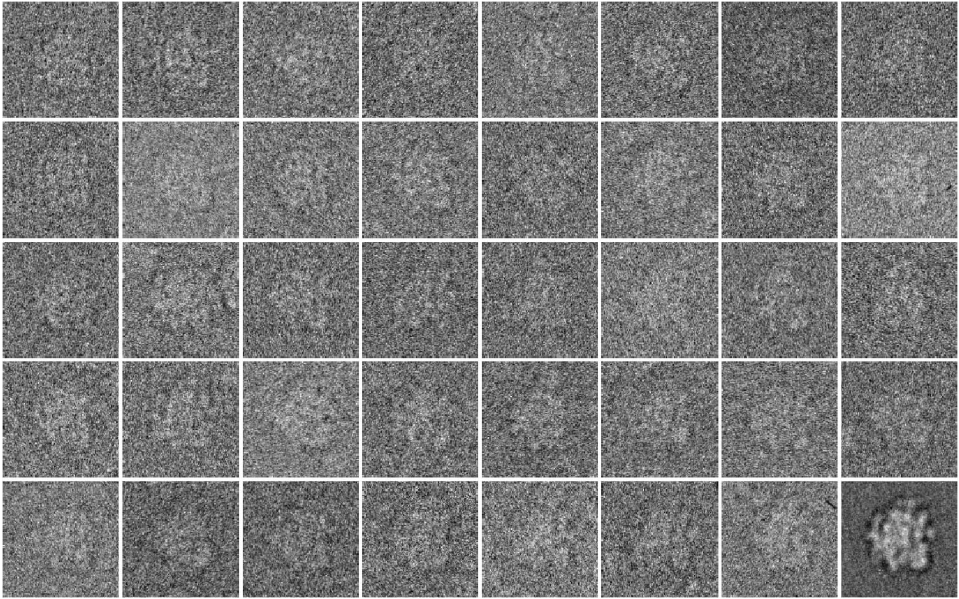
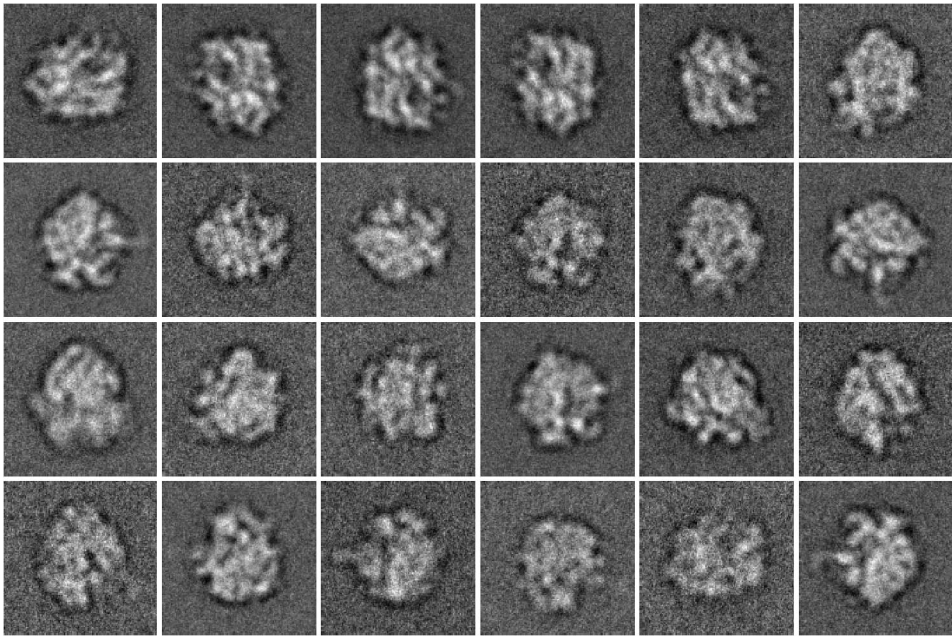


FIG. 7. One γ -SUP clustering class example contains 39 element images. Their class average is shown on the right-bottom corner.

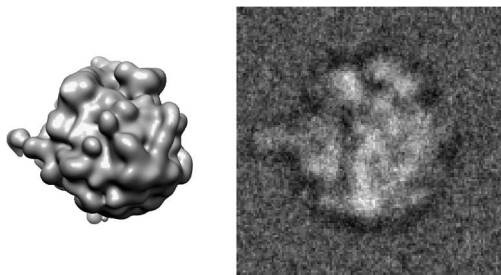
gation factor with it. The 5000 cryo-EM images were preprocessed using XMIPP [Sorzano et al. (2004)] to correct the negative phase component induced by the electron microscope transfer function. Next, the corrected images were centered and roughly aligned using the multi-reference alignment method on the SPIDER suite for single particle processing [Frank et al. (1996)]. The dimension of those images was reduced from 130×130 to 15×15 via MPCA [Hung et al. (2012)]. PCA was then further applied to the correlation matrix of those MPCA loading scores. The final dimension of each image is 20. Figure 7 shows a gallery of 39 raw cryo-EM images with a noticeably similar orientation, which were grouped into a single cluster by γ -SUP using the parameters $s = 0.025$ and $\tau = 1$, and the average of those images at the right bottom corner, which has clearly brought out an abundance of meaningful detail hidden in the raw images. This particular clustering exemplifies the potential of γ -SUP in processing real data. A total of 24 clusters with cluster sizes between 23 to 219 were found and their averages are given in Figure 8. We demonstrate their consistency with the 2D projections of known ribosome 3D structure with an example shown in Figure 9. One of the class averages is compared with a 2D projection of a 70s ribosome 3D structure solved by X-ray. The P-stalk [Wilsome and Cate (2012)] with its signature tail on the left side can be clearly seen.

6. Conclusion. We have combined a model-based q -Gaussian mixture clustering method, the γ -estimator and a self-updating process, SUP, to propose γ -

FIG. 8. 24 class averages by γ -SUP.

SUP, a novel hybrid designed to meet the image clustering challenges encountered in cryo-EM analysis. Characteristically, sets of cryo-EM images have low SNR, many of which are misaligned and should be treated as outliers, and which form a large number of clusters due to their free orientations. Because of its capability to identify outliers, γ -SUP can separate out the misaligned images and create the possibility for further correcting them. Thus, we have been able to present a successful application of γ -SUP on cryo-EM images.

Eliminating the need to set initial random cluster centers and cluster number, γ -SUP requires only the specification of (s, τ) . We have shown that γ -SUP has

FIG. 9. A single projected view of a known ribosome 3D structure compared with a single class average obtained by γ -SUP.

robust performance over s in many scenarios. Once s is chosen, the phase transition scheme may suggest a reasonable range for τ . This insensitivity with respect to s -selection and the observation of the phase transition greatly reduces the difficulty in selecting the tuning parameters. We summarize the characteristics which make for the success of γ -SUP:

- γ -SUP adopts a q -Gaussian mixture model with $q < 1$, which has compact support. Hence, it sets a finite influence range for each component and completely rejects data outside this range. When a data point is outside a certain cluster representative's influence range, it contributes zero weight toward this representative.
- γ -SUP estimates the model parameters by minimizing γ -divergence. The minimum γ -divergence downweights the influence for data that deviates far from the cluster centers, which enhances the clustering robustness.
- γ -SUP extracts clusters without the need of specifying the number of components K and random initial centers. It starts with each individual data point as a singleton cluster [i.e., with a mixture of n components, see (15)], and K is data-driven.
- γ -SUP allows singleton or extremely small-sized clusters to accommodate potential outliers.
- γ -SUP uses $\hat{F}^{(\ell)}$ to shrink the fitted mixture model toward cluster centers in each iteration. Such a shrinkage acts as if the *effective temperature* is iteratively decreasing, so that it improves the efficiency of mixture estimation.

Finally, we also remind the readers that the strength of γ -SUP is for cases where the number of clusters is large, or data are contaminated with noises/outliers, or cluster sizes are not balanced [some clusters are much bigger, or smaller, than others]. However, there is no advantage of using γ -SUP if the clustering problem arises from normal mixtures whose components have approximately similar sizes.

Acknowledgments. The authors gratefully acknowledge the Editor, the Associate Editor and two reviewers for their comments, which have substantially improved this work. The authors also acknowledge the support from the National Science Council, Taiwan.

REFERENCES

- ADRIAN, M., DUBOCHET, J., LEPAULT, J. and MCDOWALL, A. W. (1984). Cryo-electron microscopy of viruses. *Nature* **308** 32–36.
- AMARI, S.-I. and OHARA, A. (2011). Geometry of q -exponential family of probability distributions. *Entropy* **13** 1170–1185. [MR2811982](#)
- BANFIELD, J. D. and RAFTERY, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49** 803–821. [MR1243494](#)
- BASU, A., HARRIS, I. R., HJORT, N. L. and JONES, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85** 549–559. [MR1665873](#)

- BRACEWELL, R. N. (1956). Strip integration in radio astronomy. *Austral. J. Phys.* **9** 198–217. [MR0080017](#)
- CARREIRA-PERPIÑÁN, M. A. (2006). Fast nonparametric clustering with Gaussian blurring mean-shift. In *Proceeding of the 23rd International Conference on Machine Learning* 153–160. ACM, Pittsburgh, PA.
- CHANG, W.-H., CHIU, M.-K., CHEN, C.-Y., YEN, C.-F., LIN, Y.-C., WENG, Y.-P., CHANG, J.-C., WU, Y.-M., CHENG, H., FU, J. and TU, I.-P. (2010). Zernike phase plate cryo-electron microscopy facilitates single particle analysis of unstained asymmetric protein complexes. *Structure* **18** 17–27.
- CHEN, T.-L. (2013). On the convergence and consistency of the blurring mean-shift process. Available at [arXiv:1305.1040](#).
- CHEN, T.-L. and SHIU, S.-Y. (2007). A clustering algorithm by self-updating process. In *JSM Proceedings, Statistical Computing Section* 2034–2038. American Statistical Association, Salt Lake City, UT.
- CHENG, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17** 790–799.
- CICHOCKI, A. and AMARI, S.-I. (2010). Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy* **12** 1532–1568. [MR2659408](#)
- CROWTHER, R. A., AMOS, L. A., FINCH, J. T., DE ROSIER, D. J. and KLUG, A. (1970). Three dimensional reconstructions of spherical viruses by Fourier synthesis from electron micrographs. *Nature* **226** 421–425.
- DUBOCHET, J. (2012). Cryo-EM—the first thirty years. *J. Microsc.* **245** 221–224.
- EGUCHI, S., KOMORI, O. and KATO, S. (2011). Projective power entropy and maximum Tsallis entropy distributions. *Entropy* **13** 1746–1764. [MR2851127](#)
- FIELD, C. and SMITH, B. (1994). Robust estimation: A weighted maximum likelihood approach. *International Statistical Review* **62** 405–424.
- FRANK, J. (2002). Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu. Rev. Biophys. Biomol. Struct.* **31** 303–319.
- FRANK, J. (2009). Single-particle reconstruction of biological macromolecules in electron microscopy—30 years. *Q. Rev. Biophys.* **42** 139–158.
- FRANK, J. (2012). Intermediate states during mRNA-tRNA translocation. *Curr. Opin. Struct. Biol.* **22** 778–785.
- FRANK, J., RADERMACHERA, M., PENCZEKA, P., ZHUA, J., LI, Y., LADJADJ, M. and LEITHA, A. (1996). SPIDER and WEB: Processing and visualization of images in 3D electron microscopy and related fields. *Journal of Structural Biology* **116** 190–199.
- FRIGYIK, B. A., SRIVASTAVA, S. and GUPTA, M. R. (2008). Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Trans. Inform. Theory* **54** 5130–5139. [MR2589887](#)
- FUJISAWA, H. and EGUCHI, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *J. Multivariate Anal.* **99** 2053–2081. [MR2466551](#)
- FUKUNAGA, K. and HOSTETLER, L. D. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inform. Theory* **IT-21** 32–40. [MR0388638](#)
- GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. [MR2345548](#)
- GOOD, I. J. (1971). Comment on “Measuring information and uncertainty.” In *Foundation of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) 265–273. Holt, Rinehart and Winston, Toronto.
- GRASSUCCI, R., TAYLOR, D. and FRANK, J. (2011). Preparation of macromolecular complexes for cryo-electron microscopy. *Nature Protocols* **2** 3239–3246.
- HALL, R. J., NOGALES, E. and GLAESER, R. M. (2011). Accurate modeling of single-particle cryo-EM images quantitates the benefits expected from using Zernike phase contrast. *J. Struct. Biol.* **174** 468–475.

- HARTIGAN, J. A. (1975). *Clustering Algorithms*. Wiley, New York. [MR0405726](#)
- HENDERSON, R. (1995). The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q. Rev. Biophys.* **28** 171–193.
- HUNG, H., WU, P., TU, I. and HUANG, S. (2012). On multilinear principal component analysis of order-two tensors. *Biometrika* **99** 569–583. [MR2966770](#)
- JIANG, W., BAKER, M. L., JAKANA, J., WEIGELE, P. R., KING, J. and CHIU, W. (2008). Backbone structure of the infectious epsilon15 virus capsid revealed by electron cryomicroscopy. *Nature* **451** 1130–1134.
- LEPAULT, J., BOOY, F. P. and DUBOCHET, J. (1983). Electron microscopy of frozen biological suspensions. *J. Microsc.* **129** 89–102.
- LIU, H., JIN, L., KOH, S. B. S., ATANASOV, I., SCHEIN, S., WU, L. and ZHOU, Z. H. (2010). Atomic structure of human adenovirus by cryo-EM reveals interactions among protein networks. *Science* **329** 1038–1043.
- LLOYD, S. P. (1982). Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28** 129–137. [MR0651807](#)
- LU, H., PLATANIOTIS, K. N. K. and VENETSANOPOULOS, A. N. (2008). MPC: Multilinear principal component analysis of tensor objects. *IEEE Trans. Neural. Netw.* **19** 18–39.
- MANNING, C., RAGHAVAN, P. and SCHATZ, H. (2008). *Introduction to Information Retrieval*. Cambridge Univ. Press, New York.
- MCQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 291–297. Univ. California Press, Berkeley, CA.
- MOLLAH, M. N. H., SULTANA, N., MINAMI, M. and EGUCHI, S. (2010). Robust extraction of local structures by the minimum beta-divergence method. *Neural Networks* **23** 226–238.
- SAIBIL, H. R. (2000). Macromolecular structure determination by cryo-electron microscopy. *Acta Crystallographica Section D-Biological Crystallography* **56** 1215–1222.
- SHIU, S.-Y. and CHEN, T.-L. (2012). Clustering by self-updating process. Available at [arXiv:1201.1979](#).
- SINGER, A., COIFMAN, R. R., SIGWORTH, F. J., CHESTER, D. W. and SHKOLNISKY, Y. (2010). Detecting consistent common lines in cryo-EM by voting. *J. Struct. Biol.* **169** 312–322.
- SORZANO, C. O. S., MARABINI, R., VELÁZQUEZ-MURIEL, J., BILBAO-CASTRO, J. R., SCHERES, S. H. W., CARAZO, J. M. and PASCUAL-MONTANO, A. (2004). XMIPP: A new generation of an open-source image processing package for electron microscopy. *J. Struct. Biol.* **148** 194–204.
- SORZANO, C. O. S., BILBAO-CASTRO, J. R., SHKOLNISKY, Y., ALCORLO, M., MELERO, R., CAFFARENA-FERNANDEZ, G., LI, M., XU, G., MARABINI, R. and CARAZO, J. M. (2010). A clustering approach to multireference alignment of single-particle projections in electron microscopy. *Journal of Structural Biology* **171** 197–206.
- TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63** 411–423. [MR1841503](#)
- VAN HEEL, M. (1987). Angular reconstitution: A posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy* **21** 111–124.
- VAN HEEL, M., GOWEN, B., MATADEEN, R., ORLOVA, E. V., FINN, R., PAPE, T., COHEN, D., STARK, H., SCHMIDT, R., SCHATZ, M. and PATWARDHAN, A. (2000). Single-particle electron cryo-microscopy: Towards atomic resolution. *Q. Rev. Biophys.* **33** 307–369.
- WILSON, D. and CATE, J. (2012). The structure and function of the eukaryotic ribosome. *Cold Spring Harbor Perspectives in Biology* **4** a011536.
- WINDHAM, M. P. (1995). Robustifying model fitting. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57** 599–609. [MR1341326](#)

YANG, Z., FANG, J., CHITTULURU, J., ASTURIAS, F. J. and PENCZEK, P. A. (2012). Iterative stable alignment and clustering of 2D transmission electron microscope images. *Structure* **20** 237–247.

T.-L. CHEN
D.-N. HSIEH
I-P. TU
Y.-M. WU
W.-H. CHANG
S.-Y. HUANG
INSTITUTE OF STATISTICAL SCIENCE
ACADEMIA SINICA
TAIPEI, TAIWAN 11529
E-MAIL: tlchen@stat.sinica.edu.tw
dnhsieh@stat.sinica.edu.tw
iping@stat.sinica.edu.tw
marinesean@gmail.com
weihau@chem.sinica.edu.tw
syhuang@stat.sinica.edu.tw

H. HUNG
INSTITUTE OF EPIDEMIOLOGY
AND PREVENTIVE MEDICINE
NATIONAL TAIWAN UNIVERSITY
TAIPEI, TAIWAN 10055
E-MAIL: hhung@ntu.edu.tw

P.-S. WU
DEPARTMENT OF BIostatISTICS
AND BIOINFORMATICS
DUKE UNIVERSITY
DURHAM, NORTH CAROLINA 27710
USA
E-MAIL: pei.shien.wu@duke.edu