Signal Detection in Quantized Random Fields Using Dempster-Shafer Theory

Duncan Ermini Leaf, Chuanhai Liu
Department of Statistics
Purdue University
West Lafayette, IN, U.S.A.
Email: (dleaf,chuanhai)@stat.purdue.edu

***Abstract*—A new method for inferring the number of signal pixels in an image is presented. The background pixels are assumed to be independent and identically distributed Gaussian random variables quantized to an integer value. Uncertainty about the distribution of pixels in an image is represented in a Dempster-Shafer Model. In this setting, a model for the distribution of signal pixels is not necessary. The method of Maximal Belief is used to ensure that inferences behave according to a frequency interpretation. The new method is demonstrated on simulated data and compared to a Bayesian multiscale segmentation method using cryo-electron microscopic images. Application of this method to other coarse data problems is discussed in the conclusion.**

Keywords: Coarse data, cryo-electron microscopy, image processing, maximal belief, multiple testing.

I. Introduction

In this paper we present a method for inferring the number of signal pixels in an image. This is an important task in situations where a large number of images are available, but many contain only background noise. For example, a telescope may produce thousands of images while scanning the sky, but only a few contain images of stars or other interesting emission sources. By first filtering out the images with low signal pixel counts, subsequent analyses can be focused on images that are likely to contain something interesting.

To accomplish this goal, we assumed that background pixel values can be modeled as quantized Gaussian random variables. Uncertainty about the distribution of each pixel was represented as a Dempster-Shafer Model (DSM; [3], [6]) and the method of Maximal Belief (MB; [7], [5]) was applied to ensure that the resulting (p, q, r) behaved according to a credible frequency interpretation. Section II describes the credibility criteria and the application of MB to create credible DSM's. The specific DSM and inference procedure used in the present problem are discussed in section III .

Finally, the new method was tested on simulated data and several cryo-electron microscopic (cryo-EM; [4]) images. Cryo-EM methods are capable of displaying biological structures at a level of detail close to the atomic scale. However, the process is known to be very noisy and therefore serves as an interesting test case for the method presented here. For a comparison, the cryo-EM images were also segmented using a Bayesian multiscale random field method [1]. The results of these studies are discussed in section IV . Section V concludes with consideration of future applications and directions for this method.

II. Credible and Maximal Beliefs

Consider a DSM for making inferences about the mean of a Gaussian distribution from a single observation, $X \sim N(\theta, 1)$. We use the auxiliary equation, $X = \theta + \Phi^{-1}(U)$, to define a posterior DSM with focal elements indexed by an auxiliary random variable U:

$$M_X(U) = \{\theta : \theta = X - \Phi^{-1}(U)\}, \quad U \in [0, 1], \qquad (1)$$

where Φ^{-1} is the inverse function of the standard Gaussian cdf. The mass for each focal element is defined by a $Unif(0, 1)$ distribution for U. For any assertion of interest, such as $\mathcal{A} = \{\theta \leq \theta_0\}$, we have

$$p_X(\mathcal{A}) = \Pr\{M_X(U) \subseteq \mathcal{A}\},$$

$$q_X(\mathcal{A}) = \Pr\{M_X(U) \subseteq \mathcal{A}^C\}, \quad \text{and}$$

$$r_X(\mathcal{A}) = 1 - p_X(\mathcal{A}) - q_X(\mathcal{A}) = 0.$$

As discussed in [5], [7], some assertions lead to (p, q, r) that would not have a sensible frequency interpretation over repetitions of the same experiment. The following credibility criteria were proposed in [7] to give (p, q, r) a frequency interpretation:

Definition 2.1 (Credibility): Given an assertion, $\mathcal{A}$, a DSM is credible at level α for inference about $\mathcal{A}$ if

$$\forall \theta \in \mathcal{A}^C, \quad \Pr\{p_X(\mathcal{A}) \geq 1 - \alpha\} \leq \alpha, \quad \text{and} \qquad (2)$$

$$\forall \theta \in \mathcal{A}, \quad \Pr\{q_X(\mathcal{A}) \geq 1 - \alpha\} \leq \alpha. \qquad (3)$$

Essentially, (2) says that if we decide $\mathcal{A}$ is true when $p_X(\mathcal{A}) \geq 1 - \alpha$, then the probability of making the wrong decision is at most α. When $\mathcal{A}$ is actually true, (3) says that there is at most α probability of being wrong if we decide $\mathcal{A}$ is false because $q_X(\mathcal{A}) \geq 1 - \alpha$.

For some assertions, such as the point assertion $\mathcal{A} = \{\theta = \theta_0\}$, definition 2.1 is not satisfied using DSM (1). The MB method introduced by [7] expands each focal elements in the DSM to be just large enough so that the resulting (p, q, r) will be credible for any assertion of interest. To expand focal

elements, we use a random set, such as $S(U) = \{U^\star : |U^\star - 0.5| \leq |U - 0.5|\}$. See [7] and [5] for discussion of how to choose $S(U)$. Applying $S(U)$ to DSM (1) yields expanded focal elements:

$$\begin{aligned} M'_X(U) &= \{\theta : \theta = X - \Phi^{-1}(u), u \in S(U)\} \\ &= [X - \Phi^{-1}(0.5 - |U - 0.5|), \\ &\qquad X - \Phi^{-1}(0.5 + |U - 0.5|)] \end{aligned} \tag{4}$$

These expanded focal elements are still indexed by $U \in [0, 1]$ and the mass for each element is still defined by the $Unif(0, 1)$ distribution. Under the conditions discussed in [7], DSM (4) satisfies definition 2.1 for every assertion.

III. A Credible DSM for Detecting Outliers

In the present study we assumed that each pixel was quantized by rounding to the nearest integer value $\tilde{X} \in [0, ..., 255]$ (rounding up in the case of exactly 0.5 fractional part). We wished to make inference about the distribution of the underlying random variable, X, for which the quantizing process gives the interval,

$$X \in \left[\tilde{X} - 0.5, \tilde{X} + 0.5\right). \tag{5}$$

We assumed that background pixels were independent and identically distributed (i.i.d.) from $N(\mu, \sigma^2)$. The inference procedure presented here does not require any assumption about the distribution of signal pixel values. Before extending this method to interval observations, we first discuss the method developed in [8] for point observations.

A. Inference with Point Observations

Suppose that the values of the underlying Gaussian variates were exactly known instead of being in an interval. For a sample of N observations, let $(X_{(i)})_{i=1}^N$ be the order statistics and apply the transformation,

$$U_{(i)} = \Phi\left(\tfrac{X_{(i)} - \mu}{\sigma}\right). \tag{6}$$

If $\{X_i\}_{i=1}^N$ are i.i.d. samples from the $N(\mu, \sigma^2)$ distribution, then $(U_{(i)})_{i=1}^N$ is a sequence of ordered $Unif(0, 1)$ random variables. Any X_i not from $N(\mu, \sigma^2)$ can be considered an outlier.

A credible DSM for the cdf of a sequence of independent samples, $\{X_i\}_{i=1}^n$, is given in [8]. Let

$$\mathbb{S}_n^+ = \{(u_1, ..., u_n) : 0 < u_1 < ... < u_n < 1\}$$

be the space of all ordered $Unif(0, 1)$ samples of size n. To represent uncertainty about the cdf of $\{X_i\}_{i=1}^n$ i.i.d. observations, one could use $U \in \mathbb{S}_n^+$ as an auxiliary random variable to create a DSM with focal elements,

$$M_{\{X_i\}}(U) = \{F \in \mathbb{C} : (F(X_{(i)}))_{i=1}^n = U\}, \tag{7}$$

where $\mathbb{C}$ is the set of all continuous cdf's. However, this DSM does not produce credible (p, q, r) for an assertion such as $\mathcal{A} = \{F : F = F_0\}$. According to the MB method a random set can be used to expand the focal elements and make (7) credible. For $U \in \mathbb{S}_n^+$, the random set in [8] is defined as

$$S_n(U) = \{U^\star \in \mathbb{S}_n^+ : g(U^\star) \geq g(U)\},$$

where

$$g(u) = \sum_{i=1}^n \left[a_i \ln(u_i) + b_i \ln(1 - u_i)\right], \tag{8}$$

with $a_i = (n - i + 0.7)^{-1}$ and $b_i = (i - 0.3)^{-1}$. This random set always contains the point,

$$\arg\max g(U) = \left(\frac{a_1}{a_1 + b_1}, ..., \frac{a_n}{a_n + b_n}\right),$$

which is close to the point of marginal medians for ordered $Unif(0, 1)$ random variables, $(U_{(1)}, ..., U_{(n)})$. Smaller values of $g(U)$ imply a larger marginal interval for each $U_{(i)}$. Finally, as $g(U)$ approaches $-\infty$ there must be at least one $U_{(i)}$ approaching $\{0, 1\}$, and thus, $S(U)$ expands to cover the whole of $\mathbb{S}_n^+$. Applying the random set $S_n(U)$ to (7) results in the credible posterior DSM introduced in [8] with focal elements,

$$M'_{\{X_i\}}(U) = \left\{F \in \mathbb{C} : g\left((F(X_{(i)}))_{i=1}^n\right) \geq g(U)\right\}.$$

The mass for each unique focal element corresponds to the distribution of $g(U)$. Intuitively, the highest probability region of $g(U)$ corresponds to the most likely ordered $Unif(0, 1)$ samples.

To infer the number of outliers of the cdf F_0 in a sample of size N, [7] and [8] compute (p, q, r) for the sequence of assertions, $\mathcal{A}_k =$ "at least k X_i's are outliers of F_0" ($k = 1, ..., N$). For pixels distributed as $N(\mu, \sigma^2)$, $F_0(x) = \Phi(\frac{x-\mu}{\sigma})$. In this case, $\mathcal{A}_k^C =$ "there are at least $N-k+1$ X_i's from $N(\mu, \sigma^2)$." Equivalently, if we transform the ordered $X_{(i)}$'s according to (6), then $\mathcal{A}_k^C =$ "at least $N - k + 1$ $U_{(i)}$'s are ordered $Unif(0, 1)$ random variates." Formally, for $j < N$, let

$$\mathbb{U}_N^j = \{(U_{(i_1)}, ..., U_{(i_{N-j})}) : 1 \leq i_1 < ... < i_{N-j} \leq N\}$$

be the set of all length-$(N - j)$ subsequences of $(U_{(i)})_{i=1}^N$. The negated assertion can be represented in $\mathbb{S}_{N-k+1}^+$ as $\mathcal{A}_k^C = \mathbb{U}_N^{k-1}$. For $\mathcal{A}_k^C$ to be false (and $\mathcal{A}_k$ true), it would be sufficient to show that no elements in $\mathbb{U}_N^{k-1}$ are ordered $Unif(0, 1)$ variates. Of course we do not know with certainty which elements of $\mathbb{U}_N^{k-1}$ come from $Unif(0, 1)$ and which do not. Therefore, computing $p_X(\mathcal{A}_k)$ amounts to computing the belief that none of the sequences in $\mathbb{U}_N^{k-1}$ are ordered samples from $Unif(0, 1)$.

$$\begin{aligned} p_X(\mathcal{A}_k) &= \Pr\{S_{N-k+1}(U) \subseteq \mathcal{A}_k\} \\ &= \Pr\left\{\mathcal{A}_k^C \subseteq [S_{N-k+1}(U)]^C\right\} \\ &= \Pr\left\{\mathbb{U}_N^{k-1} \subseteq [S_{N-k+1}(U)]^C\right\} \\ &= \Pr\left\{\forall u \in \mathbb{U}_N^{k-1},\, g(u) \leq g(U)\right\} \\ &= \Pr\left\{\max_{u \in \mathbb{U}_N^{k-1}} g(u) \leq g(U)\right\}. \end{aligned} \tag{9}$$

Since $|\mathbb{U}_N^j| = \binom{N}{j}$ and $S_{N-k+1}(U)$ is almost surely uncountable,

$$q_X(\mathcal{A}_k) = \Pr\{S_{N-k+1}(U) \subseteq \mathcal{A}_k^C\} = 0.$$

Thus, $r_X(\mathcal{A}_k) = 1 - p_X(\mathcal{A}_k)$. An algorithm for finding the maximum of $g(u)$ in (9) is given in [8] and the probability can be estimated using Monte Carlo simulation.

B. Inference with Coarse Observations

In the present problem, where only the quantized pixel values, $\{\tilde{X}_i\}_{i=1}^N$, are observed, applying the transformation (6) gives the following interval for each $U_{(i)}$:

$$w_i = \left[\Phi\left(\frac{\tilde{X}_{(i)} - 0.5 - \mu}{\sigma}\right), \Phi\left(\frac{\tilde{X}_{(i)} + 0.5 - \mu}{\sigma}\right)\right).$$

The method of [8] can be applied here as well, even though the $U_{(i)}$ values are not precisely known. Let

$$\mathbb{W}_N^j = \{(w_{i_1}, ..., w_{i_{N-j}}) : 1 \le i_1 < ... < i_{N-j} \le N\}.$$

Then,

$$\mathcal{A}_k^C = \{U^\star \in \mathbb{S}_{N-k+1}^+ : \exists W \in \mathbb{W}_N^{k-1}, U^\star \in W\}$$

and

$$\begin{aligned} p_X(\mathcal{A}_k) &= \Pr\left\{\mathcal{A}_k^C \subseteq [S_{N-k+1}(U)]^C\right\} \\ &= \Pr\left\{\max_{W \in \mathbb{W}_N^{k-1}} \sup_{u \in W} g(u) \le g(U)\right\}, \end{aligned} \tag{10}$$

where g is defined as in (8). Note that

$$a_i \ln(v) + b_i \ln(1-v)$$

is concave in v and for each $w_j \in W$,

$$\sup_{v \in w_j} a_i \ln(v) + b_i \ln(1-v)$$

can be found analytically. Therefore,

$$\sup_{u \in W} g(u) = \sum_{j=1}^{N-k+1} \sup_{v \in w_j} [a_j \ln(v) + b_j \ln(1-v)] \tag{11}$$

can be computed analytically. The supremum of $g(u)$ over $W \in \mathbb{W}_N^{k-1}$ in (10), can be found by using the algorithm of [8] with (11) as the objective function.

In practice, to estimate the number of signal pixels in an image of size N one could use a binary search over $k = 1, ..., N$ to find $\hat{k}_\alpha = \max\{k : p_X(\mathcal{A}_k) \ge 1 - \alpha\}$. The choice of α depends on the acceptable level of risk in overestimating the number of signal pixels.

IV. Results

A. Simulation Study

First, the method of section III was used to replicate the simulation study of [7], but using coarse data. The variates, $X_i \sim N(\mu_i, 1)$ $(i = 1, ..., 100)$, were generated according to one of the following regimes:

(a) $\mu_1, ..., \mu_{100} = 0$
(b) $\mu_1, ..., \mu_{90} = 0$ and for $i = 91, ..., 100$, $\mu_i \sim 2 + Exp(1)$
(c) $\mu_1, ..., \mu_{90} = 0$ and for $i = 91, ..., 100$, $\mu_i \sim 4 + Exp(1)$
(d) $\mu_1, ..., \mu_{90} = 0$ and for $i = 91, ..., 100$, $\mu_i \sim 6 + Exp(1)$

Each observation was rounded to the nearest integer, giving the interval in (5). Ten simulations were performed for each of the four regimes. For each simulation the sequence $\{p_X(\mathcal{A}_k)\}_{k=1}^{21}$ was computed. These sequences are plotted in Fig. 1. This study was implemented in R and it took approximately 57 minutes to complete all forty simulations on a desktop PC with 3.25 GB of RAM and a 3.0 GHz processor.

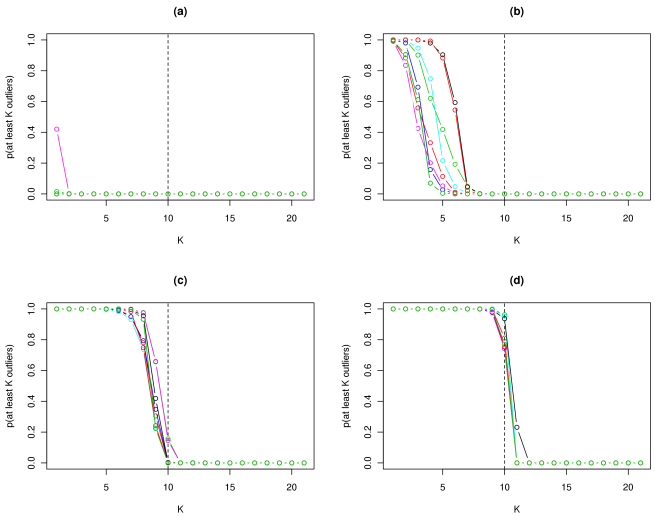

Figure 1. Replication of the study in [7], but using the method of section III and quantized data.

B. Cryo-EM Image Study

Next, the method was applied to cryo-EM images. Three images were chosen to represent high, medium, and low quantities of signal pixels. Each image was 221×216 pixels. The quantizing model in (5) was assumed. Estimates of the noise parameters, μ and σ^2, were obtained from a sample of each background region. For each image, the sequence $\{p_X(\mathcal{A}_{\lfloor qN \rfloor})\}$ was computed with $q = 0.05, 0.10, 0.15, ..., 0.80$. These sequences (up to $q = 0.5$) and the images are shown in Fig. 2. Implemented as C functions called from R, each image took approximately 3.8 hours to process on a server with 12 GB of RAM and a 2.83 GHz processor.

Additionally, for each image a binary search was performed to find $\hat{k}_{0.05}$, the estimated number of signal pixels. Because

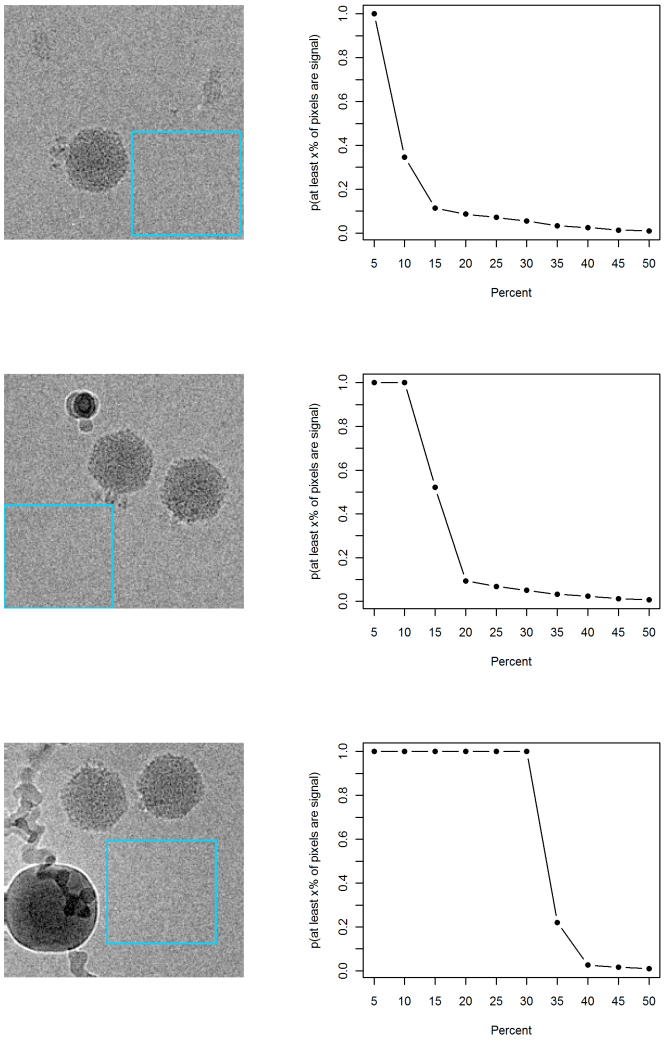

Figure 2. Results for cryo-EM images with low (top row), medium (middle row), and high (bottom row), amounts of signal pixels. The left column shows the original image with the background sample region for estimating μ and σ^2 outlined by a rectangle. The right column plots the $\{p_X(\mathcal{A}_{\lfloor qN \rfloor})\}$ sequence for $q = 0.05, 0.10, 0.15, ..., 0.50$.

a background pixel is independent of all other pixels, it is uncorrelated with its neighbors. For a given pixel, i, we defined a neighborhood, ∂i, as the 15×15 square around, but not including pixel i. We then calculated the following measure of correlation between the pixel and its neighbors:

$$s_i = \sum_{j \in \partial i} \frac{(x_i - \mu)(x_j - \mu)}{m},$$

where m is the number of pixels in ∂i. The $N - \hat{k}_{0.05}$ pixels with the lowest values of $|s_i|$ were considered background. For a comparison, we also performed two-class segmentation using the Bayesian multiscale method [1] implemented in the SMAP software package [2]. For each image, the SMAP package estimated the distribution for the background class using the previously obtained background samples while the signal class distribution was esitmated using several samples taken manually from each signal region. Fig. 3 shows each image with the estimated background pixels' values set near μ for both the present and SMAP methods.

V. Concluding Remarks

As a tool for segmentating the signal and background in cryo-EM images, the present method performed comparably to the multiscale Bayesian method [1] implemented in SMAP [2]. However, one advantage of using the present method is that it requires no information about the distribution of the signal pixels. The SMAP package, on the other hand, estimated the signal distribution using a Gaussian mixture model and training samples from the signal region. In the cryo-EM images, it was relatively easy to find a large, representative background sample. This is often more difficult for signal regions. Ultimately, any manual sampling defeats the purpose of using this method for automatically filtering out images with a low number of signal pixels. Therefore, future work should consider simultaneous estimation of the background distribution and detection of outliers from that distribution.

The method presented in this paper can be applied to any sequence of interval data, so long as all unique intervals in the sequence are disjoint. An interesting application is found in the focal elements of DSM's for discrete random variables. Consider, for example, the DSM of [3] for a single Poisson random variable, $Y \sim Poisson(\lambda)$. The posterior DSM for the rate has focal elements,

$$\lambda \in [V_Y, V_{Y+1}), \tag{12}$$

where $V_k \sim Gamma(k, 1)$. Let G_k be the $Gamma(k, 1)$ cdf. For an ordered sequence of n Poisson counts, $(Y_{(i)})_{i=1}^n$, we can apply a transformation and invert (12) to obtain an interval for an auxiliary random variable:

$$U_{(i)} \in \left[G_{Y_{(i)}}(\lambda), G_{Y_{(i)}+1}(\lambda)\right).$$

The present method can then be used to infer the number Y_i's not from the $Poisson(\lambda)$ distribution.

As noted in section III, the run time for this method was relatively long, even for small images. However, most of this time appeared to be spent on Monte Carlo estimation of (10) rather than maximization. While the algorithm in [8] lends itself to parallelization that could yield significant speedup in the maximization, future work should first consider an efficient approximation to the distribution of $g(U)$. One possibility is the distribution of $\log(\max g(u) - g(U))$, which appears to be approximately Gaussian.

References

[1] C. A. Bouman and M. Shapiro, "A multiscale random field model for Bayesian image segmentation," *IEEE Trans. Image Process.*, vol. 3, no. 2, pp. 162–177, 1994.

[2] C. A. Bouman and M. Shapiro, *SMAP 1.6.4* [Online] https://engineering.purdue.edu/∼bouman/software/segmentation/, 2005.

[3] A. P. Dempster, "The Dempster-Shafer calculus for statisticians," *Internat. J. of Approx. Reason.*, vol. 48, no. 2, pp. 365–377, 2008.

[4] R. M. Glaeser, "Cryo-electron microscopy of biological nanostructures," *Phys. Today*, vol. 61, no. 1, pp. 48–54, 2008.

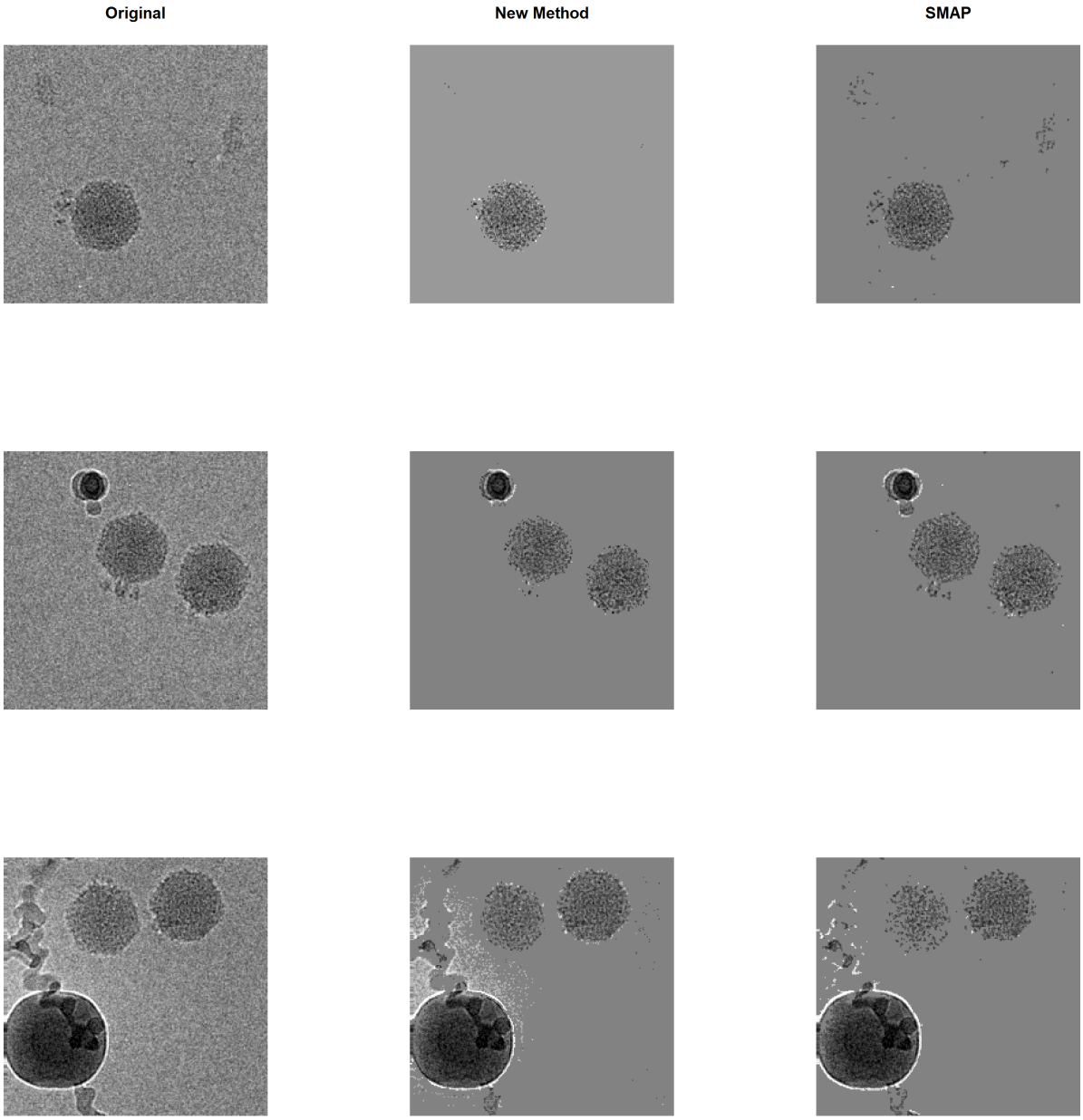

Figure 3. Results of cryo-EM image segmentation. The leftmost column shows the original image. In the middle column, the estimated background pixels using the new method with $\alpha = 0.05$ are set near the mean value. The rightmost column shows the SMAP [2] segmentaton with background pixels set near the mean value.

[5] R. G. Martin, J. Zhang, and C. Liu, "Dempster-Shafer theory and statistical inference with weak beliefs," Preprint: http://www.math.iupui.edu/∼rgmartin/Research/MZL-dswb.pdf, 2009.
[6] G. Shafer, *A mathematical theory of evidence*. Princeton Univ. Press, Princeton, NJ, 1976.
[7] J. Zhang and C. Liu, "Dempster-Shafer inference with weak beliefs," to appear in *Stat. Sinica*, Preprint: http://www.stat.purdue.edu/∼chuanhai/docs/MB.pdf, 2009.
[8] J. Zhang, J. Xie, and C. Liu, "Probabilistic inference: tests and multiple tests," Preprint: http://www.stat.purdue.edu/∼chuanhai/docs/ZXL2009.pdf, 2009.