

# Adaptive Methods for Spatial Scan Analysis via Semiparametric Mixture Models

Ramani S. PILLA, Peng TAO, and Carey E. PRIEBE

Spatial scan density (SSD) estimation via mixture models is an important problem in the field of spatial statistical analysis and has wide applications in image analysis. The “borrowed strength” density estimation (BSDE) method via mixture models enables one to estimate the local probability density function in a random field wherein potential similarities between the density functions for the subregions are exploited. This article proposes an efficient method for SSD estimation by integrating the borrowed strength technique into the alternative EM framework which combines the statistical basis of the BSDE approach with the stability and improved convergence rate of the alternative EM methods. In addition, we propose adaptive SSD estimation methods that extend the aforementioned approach by eliminating the need to find the posterior probability of membership of the component densities afresh in each subregion. Simulation results and an application to the detection and identification of man-made regions of interest in an unmanned aerial vehicle imagery experiment show that the adaptive methods significantly outperform the BSDE method. Other applications include automatic target recognition, mammographic image analysis, and minefield detection.

**Key Words:** Cyclical EM; Kernel density estimation; Missing data; Mixture distribution; Nonhomogeneity detection; Nonparametric model; Paired complete data EM algorithm; Profile likelihood; Random field; Rotated EM algorithm; Scan process; Segmentation; Semiparametric model.

## 1. INTRODUCTION

Mixture models (Titterton, Smith, and Makov 1985; Lindsay 1995; McLachlan and Peel 2001) have become one of the most widely used statistical tools in the analysis of heterogeneous data, aiding researchers in interpreting existing data or in classifying new data. Recent applications of mixture models to spatial scan analysis (Cressie 1993; Chen and Glaz

---

Ramani S. Pilla is Assistant Professor, Department of Statistics, Case Western Reserve University, Cleveland, OH 44106 (E-mail: [pilla@cwru.edu](mailto:pilla@cwru.edu)). Peng Tao is Senior Research and Development Engineer, Accu Image Diagnostics Corporation, San Francisco, CA 94080. Carey Priebe is Professor, Department of Mathematical Sciences, Johns Hopkins University, Baltimore, MD 21218 (E-mail: [cep@jhu.edu](mailto:cep@jhu.edu)).

©2003 American Statistical Association, Institute of Mathematical Statistics,  
and Interface Foundation of North America

*Journal of Computational and Graphical Statistics*, Volume 12, Number 2, Pages 332–353  
DOI: 10.1198/1061860031716

1997), particularly spatial scan density estimation (Priebe, Marchette, and Rogers 1997a; Priebe and Chen 2001), show great promise for image analysis problems (Bhanu et al. 1997; Popat and Picard 1997). For example, mammographic image analysis (Priebe 1996), automatic target recognition (ATR) (special issue on ATR of *IEEE Transactions on Image Processing* 1997; Solka et al. 1998; Priebe, Solka, and Tao 1997b) and minefield detection are special cases. One can view a spatial process as a mixture of local processes—mixing over different neighborhoods. The guiding principle behind the application of mixture models to spatial scan analysis is that one can contrast two or more regions in a scanned area by comparing their relative levels of variability captured by a mixture distribution.

Let  $\xi(x): R_0 \rightarrow \mathfrak{R}$  be a random field (Geman 1990), with a domain of definition  $R_0 \subset \mathfrak{R}^d$ . A “random field” in this context is simply an image. For simplicity, we assume that (1) the image is made up of  $r$  disjoint regions so that we can write  $R_0 = \cup_{k=1}^r R_k$  and (2) the random variable associated with each random field (or the observations from a region) are identically distributed and have the same dependence structure. The probability density function (pdf) associated with the  $k$ th random field is called the *class conditional pdf*, denoted by  $g_k(\xi)$ . That is, each  $R_k$  is a subregion of homogeneity and  $\xi(x)$ , a feature value at pixel location  $x$  in region  $R_k$ , follows the pdf  $g_k(\xi)$ . A realization of the *piecewise stationary* random field  $\xi(x)$  (Priebe 1996) is called a piecewise stationary spatial sample. For example, an image is a piecewise stationary spatial sample and the value of a field observation  $\xi(x)$  at site  $x$  represents pixel intensity. The basic idea is to classify the spatial observations adaptively into an unknown number of disjoint regions each of which is homogeneous and in turn use a simple model within the regions. However, the number, the locations, and the shapes of the regions are not known a priori. In this scenario, it is useful to obtain the information about the underlying (often unknown) class conditional pdf of the observations in each region.

The lack of knowledge of the location of differing regions necessitates that the regions obtained in the initial segmentation be small compared to the size of the true but unknown region so that the chosen region is in the interior of any of the  $R_k$ . Due to the small effective sample size from each region, one cannot obtain an accurate nonparametric estimation of the pdfs. To obtain a parametric estimate of the density, one must make certain assumptions on the form of the densities. To this end, we assume that the class conditional pdf is of the form of a finite mixture of distributions so that the goal becomes obtaining the maximum likelihood estimators (MLEs) of the parameters that characterize the mixture density. The standard and powerful computational tool for finding the MLE of the parameters in the mixture model is the well-known EM algorithm (Dempster, Laird, and Rubin 1977). Considerable work has been done to improve the (often slow) rate of convergence of the EM algorithm in various situations [see McLachlan and Krishnan (1997) and the references therein for a general discussion and for several extensions of the EM algorithm].

The principal difficulty in the application of mixture models to spatial scan analysis is that when we look afresh at an image, we have no a priori information on the locations and spatial extents of the regions and hence we introduce a regional structure on  $R_0$ . Priebe (1996) proposed the *borrowed strength density estimation* (BSDE) method via mix-

ture models to improve the segmentation of random fields. A BSD estimator exploits the similarities between the densities in different regions of the field by combining all of the observed data to find an estimate of the invariant parameters—or equivalently the component densities—in the mixtures. In turn, the component densities are fixed at the estimated values and then mixture models are fitted in each of the many small subregions. The BSDE method is an elegant solution to many image analysis problems as it produces superior detection performance. However, it is computationally expensive due to slow convergence of the EM algorithm in the BSDE setting. The slowness of the conventional EM is worse when (1) component densities are poorly separated or (2) ML estimator requires some of the mixing coefficients to be zero. Situation (1) means that the component densities are highly correlated. As in linear regression with collinearity, this means that the parameter values can be varied widely with little effect on the density, creating in turn a relatively flat likelihood function. This collinearity also has an adverse effect on the EM algorithm since the missing data become much more informative relative to the observed data (Pilla and Lindsay 2001). The goal of this research is to improve the efficiency of the BSDE method.

Our emphasis is on both performance and computational speed. The latter is becoming increasingly important as the demands on image processing systems have increased. For example, ATR in gray-scale images—specifically, the automated detection and identification of man-made *regions of interest* (ROI) in *unmanned aerial vehicle* (UAV) imagery (Solka et al. 1998)—is the main objective of many image segmentation systems. The goal of these systems is to detect, recognize, and classify *targets* (typically man-made objects, such as buildings, tanks, and aircraft) in an image. The speed and detection performance of the ATR techniques (used in both the national defense and the manufacturing industries) in image processing and analysis play an important role in making this goal a reality. For this article, we define an ROI as targets and some parts of the “background” or “clutter.”

Section 3.1 discusses a motivating example that is concerned with a high-dimensional ATR problem cast in the framework of a mixture model. In this scenario, the convergence of the conventional EM algorithm in the BSDE setting is extremely slow since some of the component densities are severely overlapping. Pilla and Lindsay (2001) proposed alternative EM methods that significantly improve the rate of convergence of the conventional EM algorithm. In the current research we propose efficient computational methods for the SSD estimation, a synthesis of the BSDE and the alternative EM methods. The resulting methods combine the statistical basis of the BSDE technique with the stability and improved convergence rate of the alternative EM methods. Furthermore, we extend the aforementioned cyclical approaches via a “dynamic adaptation rule” for the mixing coefficients which eliminates the need to find the posterior probability of membership of the component densities afresh in each subregion. We refer to the resulting approaches *Adaptive Spatial Scan Density Estimation* methods.

Forbes and Raftery (1999) proposed a Bayesian morphology method that combines the elegance of Bayesian image analysis with the speed of mathematical morphology. Their approach involves casting the marginal distribution of the pixel intensities as a finite mixture model and in turn estimating the parameters via the EM algorithm. Although both the

Bayesian morphology and the Adaptive SSD estimation methods are based on finite mixture models, the framework and the methodology of the two approaches are very different. As pointed out by Forbes and Raftery (1999), one limitation of their approach is the need to specify some insensitive conditions which may limit the applicability to simple image and noise models. However, our method is not only computationally simple but also applicable to a larger class of problems.

The rest of this article is organized as follows: Section 2 reviews the SSD estimation and the conventional EM algorithm to estimate the scan densities. Section 3 presents the paired complete data framework for the SSD estimation and embeds the Rotated EM into the SSD estimation problem. Section 4 develops adaptive SSD estimation methods. Section 5 details some of the issues that arise in implementing our methods. Section 6 presents simulation studies to evaluate the performance of one of our adaptive methods. We investigate the adaptive SSD estimation method via the ATR experiment in Section 7. Section 8 concludes with a discussion of the merits and extensions of the methods.

## 2. BACKGROUND: SPATIAL SCAN DENSITY ESTIMATION

In this section we introduce the SSD estimation and describe its relation to image processing. Recall that  $R_0 = \cup_{k=1}^r R_k$ . As there is seldom a priori knowledge of the location of the local regions of interest  $R_k$ , it is necessary to introduce a regional structure on  $R_0$ . In general, a regional structure on  $R_0$  is introduced via the *scan process* (Cressie 1993; Priebe 1996) so that  $R_0 = \cup_{s \in R_0} N_s(\delta)$ , where  $N_s(\delta)$  is a *scan window* about the spatial location  $s$  with a *size index*  $\delta$ . For a two-dimensional discrete image, one possible choice for  $N_s(\delta)$  is a  $(2\delta + 1) \times (2\delta + 1)$  square-shaped “moving scan window”:  $N_s(\delta) = \{(s_0 + p, s_1 + q) \in R_0 : -\delta \leq p, q \leq \delta\}$ , where  $s = (s_0, s_1) \in R_0$  is the location of the pixel in the image. If scan window  $N_s(\delta)$  is entirely within the region  $R_k$ , the class conditional pdf  $g_k$  is same as the scan density  $g_s$ ; otherwise they are different. The choice of the scan regions as balls or rectangles is artificial. In practice, one can use an image segmentation algorithm to partition  $R_0$  into a disjoint union of subregions which in turn can be used as the scan regions for incorporating the edge information into the process. However, this leads to an added complexity of unequal and random sample sizes. This was considered by Priebe et al. (1997a) in the mammography analysis.

In the development of the theory for the SSD estimation, one assumes that the observations are identically distributed but dependent and the stationarity assumption holds. Although neither of these assumptions are particularly realistic, the methodology seem to exhibit robustness to their violation. Future work involves developing theory that will take these into consideration.

Let  $\mathbf{y}_s = (y_1^s, \dots, y_{n_s}^s)$  be the observation vector from the scan window  $N_s(\delta)$ . The pdf of  $\mathbf{y}_s$ , called the *scan density* or *scan pdf* and denoted by  $g_s$ , plays an important role in many statistical image analysis applications such as nonhomogeneity analysis and ATR. For example, in nonhomogeneity analysis, one obtains the scan density estimate  $\hat{g}_s$  using the features from the scan window  $N_s(\delta)$  to test the hypothesis  $H_0$ : homogeneity ( $g_s = g_t \forall s, t$ )

against  $H_1$ : nonhomogeneity ( $\exists s, t \ni g_s \neq g_t$ ) via the test statistic  $T = \max_{s,t} d(\hat{g}_s, \hat{g}_t)$ , where  $d(x, y)$  denotes distance between  $x$  and  $y$ . For the reasons discussed in Section 1, we model the underlying scan densities parametrically by assuming that the  $g_s$  is a finite mixture of  $m_s$  absolutely continuous exponential family densities. That is,  $g_s = g(\mathbf{y}_s; \Psi_s) = \sum_j \pi_j^s \varphi(\mathbf{y}_s; \theta_j^s)$ , where  $\Psi_s = (\theta_s, \pi_s) = [(\theta_1^s, \dots, \theta_{m_s}^s), (\pi_1^s, \dots, \pi_{m_s}^s)]'$ ,  $\sum_j \pi_j^s = 1$  and  $\pi_j^s \geq 0$  for all  $j = 1, \dots, m_s$ . The  $s, j$  and  $i$  subscripts identify the scan window, mixture component and the observation, respectively. The component parameters  $\theta_j^s$ 's and the weights  $\pi_j^s$ 's are the *support points* and the *probability masses* of the mixing distribution, respectively. In this article, we assume that  $\varphi$  is a normal density so that  $\theta_j^s = (\mu_j^s, \sigma_j^s)$  for  $\mu_j^s \in \mathbb{R}$  and  $\sigma_j^s \in (0, \infty)$ ; however, the treatment is applicable in general.

The underlying mixture components of the scan densities are invariant across the entire field domain  $R_0$  in terms of their location in the parametric space, and hence these pdfs differ only in their mixing coefficients,  $\pi_s$ . A borrowed strength estimate exploits this invariance by using all of the observed data to find an estimate of the invariant parameters and imposing this estimate as a constraint in estimating the scan pdf. The BSDE method can be summarized as follows:

**Step 1:** On observing the overall sample  $\mathbf{y} = (y_1, \dots, y_n)$  from an image that is currently being processed, estimate the number of mixture components and the common mixture density of the features via the alternating kernel and mixture (AKM) algorithm (Priebe and Marchette 2000) as  $g(\mathbf{y}; \hat{\Psi}) = \sum_{j=1}^m \hat{\pi}_j \varphi(\mathbf{y}; \hat{\theta}_j)$ , where  $\hat{\Psi} = [(\hat{\theta}_1, \dots, \hat{\theta}_m), (\hat{\pi}_1, \dots, \hat{\pi}_m)]$ .

**Step 2:** Based on the local sample  $\mathbf{y}_s = (y_1^s, \dots, y_{n_s}^s)$  from the scan window  $N_s(\delta)$  and the estimates of  $m$  and  $\theta$  from step 1, the profile likelihood estimate of  $\pi_s$  is obtained by maximizing the regional profile likelihood

$$L(\pi_s | \hat{\theta}, \mathbf{y}_s) = \prod_{i=1}^{n_s} g(y_i^s; \pi_s | \hat{\theta}) = \prod_{i=1}^{n_s} \sum_{j=1}^m \pi_j^s \varphi(y_i^s; \hat{\theta}_j). \tag{2.1}$$

For the finite mixture case, (2.1) is the standard profile likelihood estimate of Cox and Reid (1987). The resulting scan density estimator, denoted  $\hat{g}_s = g(\mathbf{y}_s; \hat{\pi}_s | \hat{\theta})$ , is a ‘‘local density estimator’’ as it is based only on the local sample  $\mathbf{y}_s$  from the scan window  $N_s(\delta)$ . The estimator,  $(\hat{\theta}, \hat{\pi}_s)$ , is called the *borrowed strength MLE*. Priebe (1996) showed that the BSDE yields superior results compared to the conventional local likelihood methods. The conventional local likelihood method is based on maximizing the regional local likelihood function  $L(\Psi_s | \mathbf{y}_s) = \prod_i g(y_i^s; \theta_s, \pi_s)$  to obtain the joint maximum likelihood estimators of  $\theta_s$  and  $\pi_s$ . For the subsequent spatial scan analysis, we fix the number of components and the common support at  $m$  and  $\hat{\theta}$ , respectively, for the model in (2.1). From here onwards, we use a shorter notation  $\varphi_j(y_i^s)$  for  $\varphi(y_i^s; \hat{\theta}_j)$  as our interest lies only in estimating the mixing coefficients. For many practical applications, the time critical part of the algorithm is in obtaining the profile estimator of the mixing coefficient vector  $\pi_s$  for each scan window, as this step must be performed in real-time. Thus the number of times the profile likelihood estimation criterion is evaluated is relevant.

**Remark 1.** Note that the pdf of  $\mathbf{y}_s$ , the observation vector from the scan window

$N_s(\delta)$ , denoted by  $g_s$  is a mixture density and should not be confused with an  $n_s$ -dimensional density. The subscript “ $s$ ” identifies the window from which the observation vector is drawn.

**Remark 2.** The AKM algorithm proceeds by alternating between the filtered kernel estimation (FKE) proposed by Marchette et al. (1996) and the finite mixture estimation methods. The AKM finds an estimate of the number of mixture terms  $m$ , by increasing the complexity of the mixture model based on the mismatch between the current  $m$ -component mixture and the FKE. The latter is based on a multiple-bandwidth kernel estimation and data-driven smoothing with an added advantage of not requiring a critical choice of the bandwidth parameter.

**Maximization of the Regional Profile Likelihood:** For a given  $\hat{\theta}$ , one can maximize (2.1) via the conventional EM algorithm to obtain the estimator of the weight vector  $\pi_s$ . On observing the  $\mathbf{y}_s$  from the scan window  $N_s(\delta)$  one can construct the hypothetical “complete” data as  $\mathbf{x}_s = (\mathbf{y}_s, \mathbf{z}_s)$ , where  $\mathbf{z}_s = (\mathbf{z}_1^s, \dots, \mathbf{z}_{n_s}^s)$  and  $\mathbf{z}_i^s = (z_{i1}^s, \dots, z_{im}^s)'$  with  $z_{ij}^s = 1$  or 0 depending on whether  $y_i^s \in N_s$  has been drawn from the  $j$ th component density or not. Note that  $\mathbf{z}_i^s \sim \text{Mult}(m, \pi_s)$ . At the  $(c + 1)$ st step, the conventional EM algorithm assigns mass  $(\pi_j^s)^{c+1} = n_s^{-1} \sum_i \tilde{z}_{ij}$  to  $\hat{\theta}_j$ , where

$$\tilde{z}_{ij}^s = E[z_{ij}^s | \mathbf{y}_s, \pi_s^c] = \frac{(\pi_j^s)^c \varphi_j(y_i^s)}{[\sum_j (\pi_j^s)^c \varphi_j(y_i^s)]} \tag{2.2}$$

is the posterior probability that the observation  $y_i^s$  is drawn from the  $j$ th component for  $i = 1, \dots, n_s; j = 1, \dots, m$ . See McLachlan and Krishnan (1997) for more details. From here onwards, we will refer to the BSDE via the conventional EM algorithm as the *Traditional BSDE* method.

### 3. PAIRED COMPLETE DATA FRAMEWORK FOR THE SSD ESTIMATION PROBLEM

In this section, first we demonstrate the slow convergence of the conventional EM algorithm through an application of the SSD estimation method to an ATR problem. Next, we will set up the SSD estimation problem in the paired complete data frame work of Pilla and Lindsay (2001). As illustrated in Section 1, the conventional EM algorithm is particularly slow when the (1) component densities are poorly separated or (2) mixing coefficients are on the boundary of the parametric space.

#### 3.1 EXAMPLE: FINITE MIXTURE MODEL

The ML classifier requires that the user knows a priori the parameters that define the density for the image under consideration. Alternatively, representative “trial data” must be available, consisting of pixels whose class is known. Figure 1(a) shows one of the images collected by Solka et al. (1998) which were extracted from a video tape of a UAV test flight over the Naval Strike Warfare Center, Fallon, Nevada, in the summer of 1995. For this article, we use class to mean a target (a tank or a building), clutter, or background. A training image provides information about the regions  $R_k$  and the corresponding pdf

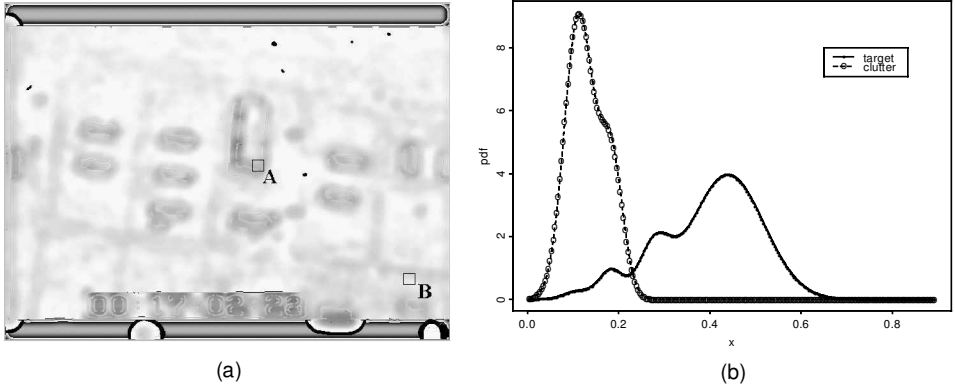


Figure 1. An example of a mixture model with poorly separated components and zero mixing coefficients: (a) The feature image used for training (target and clutter regions are annotated as “A” and “B”, respectively) and (b) the class conditional density estimates  $\hat{g}_1$  and  $\hat{g}_2$ .

estimates  $\hat{g}_k$  for all  $k$  (see Priebe 1996, figs. 2(f) and 2(g)). On observing the representative feature values  $\mathbf{y}$ , called the *training data*, from the gray-level image in Figure 1(a), a subset of these pixels were pre-identified as belonging to class 1 (or target) or class 2 (or clutter). In essence, training data  $\mathbf{y} = \{y_k(x_1), \dots, y_k(x_{n_k})\}$  representing a value at the pixel locations  $x$  in region  $R_k$  for which the true class is known is used to build the model. The resulting model is used in assessing the accuracy of the estimate obtained from the model at hand. The marginal distribution of gray-scale pixel intensities is assumed to be in the form of finite mixture of distributions. The AKM estimation method was used to find the four-component segmentation and the estimate of the common support:  $\hat{\theta} = (\hat{\mu}_1 = 0.1124, \hat{\sigma}_1 = 0.000984; \hat{\mu}_2 = 0.1832, \hat{\sigma}_2 = 0.000662; \hat{\mu}_3 = 0.2790, \hat{\sigma}_3 = 0.001187; \hat{\mu}_4 = 0.4397, \hat{\sigma}_4 = 0.006205)$ . These four-component densities per class are sufficient to provide reasonably good density estimators.

Let  $g_1$  and  $g_2$  be the normal mixture densities corresponding to the scan windows  $N_1(\delta)$  (associated with target region) and  $N_2(\delta)$  (associated with clutter region), respectively, with  $\delta = 8$ . Thus, each window is of size  $17 \times 17$  with  $(2\delta + 1)^2 = 289$  observations. For a given  $\hat{\theta}$ , the MLE of the scan densities are obtained via the conventional EM algorithm as  $\hat{g}_1 = g(\mathbf{y}_1; \hat{\pi}_1 | \hat{\theta}) = 0.022 \varphi_1(\mathbf{y}_1) + 0.058 \varphi_2(\mathbf{y}_1) + 0.134 \varphi_3(\mathbf{y}_1) + 0.786 \varphi_4(\mathbf{y}_1)$  and  $\hat{g}_2 = g(\mathbf{y}_2; \hat{\pi}_2 | \hat{\theta}) = 0.709 \varphi_1(\mathbf{y}_2) + 0.291 \varphi_2(\mathbf{y}_2)$ , respectively.

Figure 1(b) shows the scan density estimates for the target and clutter pixels, respectively. It is clear from Figure 1(b) that the component densities within each class are not only poorly separated but the mass associated with the third and fourth component densities corresponding to the clutter pixels are zeros [see Priebe (1996) and Popat and Picard (1997) for more such examples]. Recall that the  $j$ th and  $k$ th component densities in a scan window  $s$  are well separated in the sense that  $\varphi_j(\mathbf{y}_s)/g(\mathbf{y}_s; \pi_s | \hat{\theta}) \cdot \varphi_k(\mathbf{y}_s)/g(\mathbf{y}_s; \pi_s | \hat{\theta}) \approx 0$  and poorly separated in the sense that  $\varphi_j(\mathbf{y}_s)/g(\mathbf{y}_s; \pi_s | \hat{\theta}) \approx \varphi_k(\mathbf{y}_s)/g(\mathbf{y}_s; \pi_s | \hat{\theta})$  for  $j \neq k$ . This example demonstrates the extremely slow convergence of the conventional EM. See Titterton et al. (1985) and Pilla and Lindsay (1996, 2001) for more details on why the conventional EM performs poorly in these scenarios.

### 3.2 MODEL SETUP

From Section 3.1, it is clear that our interest lies only in estimating the  $\pi$  and hence we treat the component densities  $\varphi_j(\mathbf{y}_s)$  as known. The paired complete data (PCD) framework of Pilla and Lindsay (2001) starts with the choice of a pairing of densities, say  $[\{\varphi_1(\mathbf{y}_s), \varphi_2(\mathbf{y}_s)\}, \dots, \{\varphi_{m-1}(\mathbf{y}_s), \varphi_m(\mathbf{y}_s)\}]$  such that the pairing reflects correlated pairs; the last density  $\varphi_m(\mathbf{y}_s)$  is treated separately if  $m$  is odd. For the rest of the article, we assume  $m$ , the total number of densities used in  $g_s$ , is even and the nearby component densities are highly correlated. Continuing with the notation in Section 2, we define  $\mathcal{Z}_{i1}^s = (z_{i1}^s + z_{i2}^s), \dots, \mathcal{Z}_{im/2}^s = (z_{i(m-1)}^s + z_{im}^s)$  with  $(z_{ij}^s + z_{i(j+1)}^s) = 1$  or  $0$  according to whether  $y_i^s$  belongs to the  $[\varphi_j, \varphi_{j+1}]$ th pair of densities or not. The *paired complete data* becomes:  $\mathbf{x}_s^* = (\mathbf{y}_s, \mathcal{Z}_s)$ , where  $\mathcal{Z}_s = (\mathcal{Z}_1^s, \dots, \mathcal{Z}_{m/2}^s)$  and  $\mathcal{Z}_k^s = (\mathcal{Z}_{1k}^s, \dots, \mathcal{Z}_{n_s k}^s)$  for all  $k = 1, \dots, m/2$ .

The mixture density  $g(y_i^s; \boldsymbol{\pi}_s | \hat{\boldsymbol{\theta}}) = \sum_{j=1}^m \pi_j^s \varphi_j(y_i^s)$  can now be reparameterized as

$$\sum_{k=1}^{m/2} \rho_k^s [\alpha_{2k-1}^s \varphi_{2k-1}(y_i^s) + \alpha_{2k}^s \varphi_{2k}(y_i^s)],$$

where  $\rho_k^s = (\pi_{2k-1}^s + \pi_{2k}^s)$ ,  $\alpha_{2k-1}^s = \pi_{2k-1}^s / (\pi_{2k-1}^s + \pi_{2k}^s)$  and  $\alpha_{2k}^s = (1 - \alpha_{2k-1}^s)$ . Let  $\boldsymbol{\rho}_s = (\rho_1^s, \dots, \rho_{m/2}^s)$  and  $\boldsymbol{\alpha}_s = (\alpha_1^s, \dots, \alpha_m^s)$  for all  $k = 1, \dots, m/2$ . The PCD likelihood function

$$L_{PCD}(\boldsymbol{\pi}_s | \hat{\boldsymbol{\theta}}, \mathbf{x}_s^*) = \prod_{i=1}^{n_s} [\pi_1^s \varphi_1(y_i^s) + \pi_2^s \varphi_2(y_i^s)]^{\mathcal{Z}_{i1}^s} \dots [\pi_{m-1}^s \varphi_{m-1}(y_i^s) + \pi_m^s \varphi_m(y_i^s)]^{\mathcal{Z}_{im/2}^s}$$

becomes

$$L_{PCD}(\boldsymbol{\rho}_s, \boldsymbol{\alpha}_s | \hat{\boldsymbol{\theta}}, \mathbf{x}_s^*) = \prod_i \prod_k [\rho_k^s]^{\mathcal{Z}_{ik}^s} [\alpha_{2k-1}^s \varphi_{2k-1}(y_i^s) + \alpha_{2k}^s \varphi_{2k}(y_i^s)]^{\mathcal{Z}_{ik}^s}. \tag{3.1}$$

Note that  $\sum_k \rho_k^s = 1$  and  $(\alpha_{2k-1}^s + \alpha_{2k}^s) = 1$  for all  $k = 1, \dots, m/2$ . Thus the  $\alpha$  parameters act as relative weights of the pairs. From the above formulation, it is clear that the parameter vectors  $\boldsymbol{\rho}_s$  and  $\boldsymbol{\alpha}_s$  separate out in the complete data log likelihood and log  $L_{PCD}$  can be written as a sum of  $\mathcal{A}(\boldsymbol{\rho}_s) = \sum_i \sum_k \mathcal{Z}_{ik}^s \log \rho_k^s$  and  $\sum_k \mathcal{B}(\alpha_{2k}^s) = \sum_k \sum_i \mathcal{Z}_{ik}^s \log [\alpha_{2k-1}^s \varphi_{2k-1}(y_i^s) + \alpha_{2k}^s \varphi_{2k}(y_i^s)]$  for  $k = 1, \dots, m/2$ , where each piece depends on a single parameter.

### 3.3 SPATIAL SCAN DENSITY ESTIMATION VIA THE PAIRED EM

Following Pilla and Lindsay (2001), starting with the current values of the parameters  $\rho_k^c$  and  $\alpha_k^c$ , the E-step replaces  $\mathcal{Z}_{ik}^s$  in the  $\mathcal{A}$  and  $\mathcal{B}$  functions with its conditional expectation given by

$$\tilde{\mathcal{Z}}_{ik}^s = (\rho_k^c)^c \frac{\{(\alpha_{2k-1}^c)^c \varphi_{2k-1}(y_i^s) + (\alpha_{2k}^c)^c \varphi_{2k}(y_i^s)\}}{\sum_k (\rho_k^c)^c \{(\alpha_{2k-1}^c)^c \varphi_{2k-1}(y_i^s) + (\alpha_{2k}^c)^c \varphi_{2k}(y_i^s)\}} \tag{3.2}$$



which is the posterior probability that  $y_i^s$  is drawn from the  $(\varphi_{2k-1}^s, \varphi_{2k}^s)$ th pair of densities. Due to the separated parameters in the log-likelihood, the optimization problem in the M-step is simplified and one can maximize  $\mathcal{A}(\rho_s)$  explicitly, subject to the constraint  $\sum_k \rho_k^s = 1$ . However, one cannot maximize the univariate function  $\mathcal{B}(\cdot)$  explicitly. Following Pilla and Lindsay, we apply a single Newton–Raphson (NR) on each of the  $\mathcal{B}$  functions in the M-step which turns out to be highly effective, as the NR for the univariate parameter  $\alpha_{2k-1}^s$  is not only quadratically convergent but also is nearly monotonic (Böhning and Lindsay 1988). The *Paired EM* iterates between the E- and the M-steps until convergence. The original mixing coefficients can be retrieved using  $\pi_{2k-1} = \rho_k \alpha_{2k-1}$  and  $\pi_{2k} = \rho_k \alpha_{2k}$  for all  $k = 1, \dots, m/2$ .

The rate of convergence of an EM algorithm is independent of the number of NR-steps within each EM-step and hence one NR-step is sufficient (Lange 1995; proposition 1). See Pilla and Lindsay (2001) for details on practical implementation.

### 3.4 SPATIAL SCAN DENSITY ESTIMATION VIA THE ROTATED EM ALGORITHM

The Paired EM stays with one fixed pairing of densities in each EM-step. Pilla and Lindsay (2001) considered changing the missing data formulation between the EM-steps to achieve improved acceleration in all directions of the parameter space. We present their simplest approach, which involves *rotating* through a sequence of different pairings of densities. Suppose one starts with six densities such that the adjacent pairs are most similar. One possible choice of a rotation cycle involves alternating between the pairing  $(\varphi_1^s, \varphi_2^s)$ ,  $(\varphi_3^s, \varphi_4^s)$  and  $(\varphi_5^s, \varphi_6^s)$  in an odd EM-step and  $(\varphi_2^s, \varphi_3^s)$ ,  $(\varphi_4^s, \varphi_5^s)$  and  $(\varphi_6^s, \varphi_1^s)$  in an even EM-step. We apply the Paired EM within each pairing scheme. By changing the missing data formulation between the EM-steps one is able to accelerate a complementary set of relative weights, namely the  $\alpha$  parameters. To make effective use of the Paired EM, we construct the pairing schemes such that the paired densities are similar.

One could consider other cyclical algorithms, within the BSDE framework, developed by Pilla and Lindsay (2001); however, they found that this simplest method provides major gains in the rate of convergence and hence we use this approach in our article. They termed this EM based on the PCD augmentation with the above two step rotation cycles as the *Rotated EM*. Pilla (1997) showed that the Rotated EM has a better asymptotic rate of convergence than that of the conventional EM. We will demonstrate the effectiveness of this Rotated EM in the BSDE framework, particularly when the solution  $\hat{\pi}_s$  has many zeros, through our numerical studies.

## 4. AN ADAPTIVE SSD ESTIMATION METHOD BASED ON ROTATED EM ALGORITHM

In applications of SSD estimation to image analysis, a  $(2\delta + 1) \times (2\delta + 1)$ -pixel moving scan window is scanned throughout the region. At each site, the scan pdf is estimated via the BSDE method. In general, to improve the detection accuracy one creates the scan windows such that there is a substantial overlapping region between any two adjacent

ones. Overlapping of two scan windows  $N_s$  and  $N_t$ , say, implies that the mixture densities corresponding to these windows are “similar” in the sense that  $d(\pi_s, \pi_t) < \text{tol}$ , where  $d(\cdot, \cdot)$  is a distance measure (under a specified norm) and  $\text{tol}$  is a prespecified tolerance. Equivalently, in the complete data framework for the mixture model, there is a substantial overlapping of both the observed and the missing data. In this scenario, it is wasteful to replace the entire missing data in each scan window by their conditional expectations—that is, finding the posterior probability of membership of the component densities afresh in each subregion. This motivates our *dynamic adaptation* of the mixing coefficients and in turn leading to one version of the adaptive SSD estimation methods.

### 4.1 DYNAMIC ADAPTATION OF THE MIXING COEFFICIENTS

Let us assume that the  $(2\delta + 1) \times (2\delta + 1)$  scan window at site  $s$  is moving to a neighboring site  $t$ . Let  $N_s(\delta)$  and  $N_t(\delta)$  (denoted  $N_s$  and  $N_t$ , respectively, for exposition) be two overlapping scan windows of size  $\delta$  centered at  $s$  and  $t$ , respectively, such that  $N_s$  is processed before  $N_t$ . Let us assume that the cardinality of  $N_s$  or  $N_t$  be  $n$  and the cardinality of the difference set  $(N_s - N_t)$  or  $(N_t - N_s)$  be  $n_1$ . We also assume that  $n$  is much greater than  $n_1$  so that there is a substantial overlapping region between the two windows. We decompose  $N_s$  and  $N_t$  such that  $N_s = (N_s \cap N_t) \cup (N_s - N_t)$ ,  $N_t = (N_s \cap N_t) \cup (N_t - N_s)$ , where  $(N_s - N_t) = \{u | u \in N_s, u \notin N_t\}$ . Suppose the observed values in  $N_s$  and  $N_t$  are  $\{y_1, \dots, y_{n_1}, \dots, y_n\}$  and  $\{y_{n_1+1}, \dots, y_n, \dots, y_{n_2}\}$ , respectively, so that the pixels  $\{1, \dots, n_1, \dots, n\} \in N_s$  and  $\{n_1 + 1, \dots, n, \dots, n_2\} \in N_t$ . Note that  $n_2 = (n_1 + n)$  since  $N_s$  and  $N_t$  are of the same size, namely  $n = (2\delta + 1)^2$ . We have  $\{1, \dots, n_1\} \in (N_s - N_t)$ ,  $\{n + 1, \dots, n_2\} \in (N_t - N_s)$  and  $\{n_1 + 1, \dots, n\} \in N_s \cap N_t$ .

Suppose the center of the scan window is moved from a site  $s$  to a neighboring site  $t$ , then the pixels  $\{1, \dots, n_1\}$  will be replaced with that of  $\{n + 1, \dots, n_2\}$  in  $N_t$ . In general, the larger the adjacent scan window, the higher the number of overlapping pixels. Suppose a  $5 \times 5$  scan window is used in the case of a two-dimensional image. Let  $s = (s_0, s_1)$  and  $t = (s_0 + 1, s_1)$  so that  $N_s \cap N_t = \{(s_0 + p, s_1 + q) | p = -1, 0, 1, 2; q = -2, -1, 0, 1, 2\}$ ,  $(N_s - N_t) = \{(s_0 - 2, s_1 + q) | q = -2, -1, 0, 1, 2\}$  and  $(N_t - N_s) = \{(s_0 + 3, s_1 + q) | q = -2, -1, 0, 1, 2\}$ . It is clear that  $N_s \cap N_t$  containing 20 pixels overlaps with  $N_t$  and only five old pixels from  $(N_s - N_t)$  were replaced by five new ones from  $(N_t - N_s)$ .

Let us assume that the scan window  $N_s(\delta)$  is processed first and in turn the local mixture density estimate  $\hat{g}_s = g(\mathbf{y}_s; \hat{\pi}_s | \hat{\theta}) = \sum_{j=1}^m \hat{\pi}_j^s \varphi_j(\mathbf{y}_s)$  that characterizes the features of this window is obtained. In what follows, we present an efficient method to estimate the mixture density  $g_t = g(\mathbf{y}_t; \pi_t | \hat{\theta}) = \sum_{j=1}^m \pi_j^t \varphi_j(\mathbf{y}_t)$  that characterizes the features of the scan window  $N_t$  which is to be processed next. Following the notation of Section 2, we use the superscripts  $(s/t)$  and  $(s, t)$  to represent  $(N_s - N_t)$  and  $N_s \cap N_t$ , respectively. From Equation (2.2), it follows that

$$\tilde{z}_{ij}^{s/t} = \frac{(\pi_j^s)^c \varphi_j(y_i^{s/t})}{[\sum_j (\pi_j^s)^c \varphi_j(y_i^{s/t})]} \tag{4.1}$$

and

$$\tilde{z}_{ij}^{s,t} = \frac{(\pi_j^s)^c \varphi_j(y_i^{s,t})}{[\sum_j (\pi_j^s)^c \varphi_j(y_i^{s,t})]} \quad (4.2)$$

represent the posterior probabilities that  $y_i^{s/t} \in (N_s - N_t)$  and  $y_i^{s,t} \in (N_s \cap N_t)$ , respectively, are drawn from the  $j$ th component density for  $j = 1, \dots, m$ . Since  $N_s = (N_s - N_t) \cup (N_s \cap N_t)$ , we have  $\sum_{i=1}^n \tilde{z}_{ij}^s = [\sum_{i=1}^{n_1} \tilde{z}_{ij}^{s/t} + \sum_{i=n_1+1}^n \tilde{z}_{ij}^{s,t}]$ .

**Step 1:** The estimator of  $\pi_j^s$ , corresponding to the scan window  $N_s$ , obtained via the conventional EM is given by

$$\hat{\pi}_j^s = \frac{1}{n} \sum_{i=1}^n \tilde{z}_{ij}^s = \frac{1}{n} \left[ \sum_{i=1}^{n_1} \tilde{z}_{ij}^{s/t} + \sum_{i=n_1+1}^n \tilde{z}_{ij}^{s,t} \right].$$

Equivalently, we have

$$\hat{\pi}_j^{s,t} = \frac{1}{(n - n_1)} \sum_{i=n_1+1}^n \tilde{z}_{ij}^{s,t} = \frac{n}{(n - n_1)} \left[ \hat{\pi}_j^s - \frac{1}{n} \sum_{i=1}^{n_1} \tilde{z}_{ij}^{s/t} \right],$$

where  $\hat{\pi}^{s,t} = (\hat{\pi}_1^{s,t}, \dots, \hat{\pi}_m^{s,t})$  represents the coefficient vector corresponding to the region  $(N_s \cap N_t)$ .

**Step 2:** The initial value for the vector of mixing coefficients that characterizes the mixture density  $g_t$  corresponding to the scan window  $N_t(\delta)$  can be obtained as

$$\begin{aligned} (\pi_j^t)^0 &= \frac{1}{n} \sum_{i=n_1+1}^{n_2} \tilde{z}_{ij}^t = \frac{1}{n} \left[ \sum_{i=n_1+1}^n \tilde{z}_{ij}^{s,t} + \sum_{i=n+1}^{n_2} \tilde{z}_{ij}^{t/s} \right] \\ &= \hat{\pi}_j^s - \frac{1}{n} \sum_{i=1}^{n_1} \tilde{z}_{ij}^{s/t} + \frac{1}{n} \sum_{i=n+1}^{n_2} \tilde{z}_{ij}^{t/s} = \frac{(n - n_1)}{n} \hat{\pi}_j^{s,t} + \frac{1}{n} \sum_{i=n+1}^{n_2} \tilde{z}_{ij}^{t/s} \end{aligned}$$

for all  $j = 1, \dots, m$ . Since  $n \gg n_1$ , the last equation becomes

$$(\pi_j^t)^0 = \hat{\pi}_j^{s,t} + \frac{1}{n} \sum_{i=n+1}^{n_2} \tilde{z}_{ij}^{t/s}, \quad (4.3)$$

where  $\tilde{z}_{ij}^{t/s} = \hat{\pi}_j^s \varphi_j(y_i^{t/s}) / [\sum_j \hat{\pi}_j^s \varphi_j(y_i^{t/s})]$ . Thus, the starting point for  $\pi_j^t$  can be obtained by adding the contributing share of pixels  $\{n + 1, \dots, n_2\} \in N_t - N_s$  to  $\hat{\pi}_j^{s,t}$ .

The above dynamic adaptation rule takes advantage of the existing local mixture density estimator characterizing the features of the scan window  $N_s$  and provides reasonably good starting points for the mixing coefficients characterizing the mixture density corresponding to the scan window  $N_t$ . As will be shown in Sections 6 and 7, the dynamic adaptation rule simplifies the computational burden significantly. In addition, this dynamic adaptation rule can be generalized to other problems (i.e., beyond finite mixture model) that can be cast in the missing data framework (see Section 8).

## 4.2 AN OVERVIEW OF THE ADAPTIVE SSD ESTIMATION METHODS

In what follows, we summarize the *Adaptive SSD Estimation* method based on the Rotated EM algorithm.

1. On observing the overall sample  $\mathbf{y} = (y_1, \dots, y_n)$ , we estimate the common support set  $\theta$  and the number of mixture terms  $m$  via the AKM algorithm as described in Section 2.
2. Partition the domain of the piecewise stationary random field into scan windows:  $R_0 = \cup_{s \in R_0} N_s(\delta)$  for a pre-specified size index  $\delta$  as shown in Section 2.
3. For a given  $\hat{\theta}$ , obtain the estimator of the vector of mixing coefficients  $\pi_s$  corresponding to the scan window  $N_s$  via the Rotated EM algorithm illustrated in Section 3.4. Note that one can replace the Rotated EM with any of the cyclical EM algorithms of Pilla and Lindsay (2001) to create other adaptive SSD estimation methods.
4. For a given  $\hat{\pi}_s$ , update the mixing coefficients  $\pi_t$  corresponding to a neighboring scan window  $N_t$  via the dynamic adaptation rule given by Equation (4.3). Recall that the scan window is moved from a site  $s$  to a neighboring site  $t$  in such a way that there is a substantial overlapping. The dynamic adaptation rule takes advantages of this overlapping region and eliminates the need to find the estimate of the mixing coefficients afresh in each scan window.
5. Scan the window  $N_s(\delta)$  for each  $s \in R_0$  throughout the entire region  $R_0$  by repeating Steps 3 and 4.

## 5. IMPLEMENTATION ISSUES

This section addresses some of the issues that arise in the implementation of the algorithms. Following Pilla and Lindsay (2001), we compare the methods in terms of the *number of parameter updates* that occur instead of the “number of iterations.” On observing the data  $\mathbf{y}_s$  from the scan window  $N_s(\delta)$ , we find the average number of parameter updates as well as the average CPU time, where the average is over the total number of scan windows, required by each estimation method to converge to the specified tolerance level  $\varepsilon$ . Note that the CPU time is approximately proportional to the number of pixels in each scan window, namely  $(2\delta + 1)^2$ , and also affected by the tolerance parameter  $\varepsilon$ . Our simulations and experiments were run on a SUN Ultra station 1 with 128MB RAM and 143MHz clock speed.

**Convergence Criterion:** The AKM algorithm is used to estimate the number of mixture terms, the common support  $\theta$  and the vector of mixing coefficients  $\pi$ . For this step, we used the likelihood based convergence criterion: stop the algorithm if  $\log L(\theta^c, \pi^c | \mathbf{y}) - \log L(\theta^{c-10}, \pi^{c-10} | \mathbf{y}) \leq 0.005$ . Recall that the ML calculations are done only once for the entire image and hence a strict likelihood-based convergence criterion is not computationally demanding.

We need a measure of accuracy to compare the two methods. In practice, one does

not know the final value of the regional profile log-likelihood in Equation (2.1). Following Pilla and Lindsay (2001), we used the following gradient based stopping criterion because it has a solid theoretical foundation (Lindsay 1995). The gradient function creates a natural stopping rule for iterative algorithms in the mixture problem when the final log-likelihood is unknown. We stop the iterative process when  $\sup_j D_{\boldsymbol{\pi}}(j) \leq \varepsilon$ , where the gradient function is given by

$$D_{\boldsymbol{\pi}}(j) = \sum_{i=1}^{n_s} \frac{\varphi_j(y_i^s)}{g(y_i^s; \boldsymbol{\pi}^s | \hat{\boldsymbol{\theta}})} - n_s \quad \text{for all } j = 1, \dots, m. \quad (5.1)$$

It follows from Lindsay (1995, pp. 131–132) that if we set  $\varepsilon = 0.005$ , then we automatically satisfy the *ideal stopping criterion*:  $|\log L_{\text{obs}}(\hat{\boldsymbol{\pi}} | \mathbf{y}) - \log L_{\text{obs}}(\boldsymbol{\pi}^p | \mathbf{y})| \leq \varepsilon$ . This is an important measure of convergence of an ML algorithm as it provides information about the accuracy of the parameter estimates on a confidence interval scale.

## 6. SIMULATION RESULTS

This section presents a simulation experiment that evaluates the performance of the adaptive SSD estimation method for various window sizes  $\delta$  and levels of tolerance parameters  $\varepsilon$  required for convergence. In general, one has two types of images for feature recognition: *training images* (or preidentified spatial samples) and *test images* (or null segmented spatial samples). The training image consists of pixels that are preidentified—that is, the stationary random field from which the pixel comes from is known. One obtains the class conditional pdf estimates via supervised learning. The *training model*, which is a collection of the class conditional pdf estimates (Priebe et al. 1997a; Tao 2000) is built using the user identified pixels from the training images. The test image consists of pixels that are not pre-identified. The goal of an image recognition is to assign the class membership to the pixels in the test image by using the information contained in the training model.

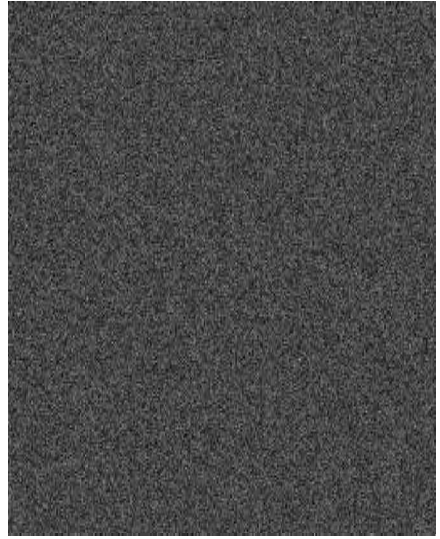
We consider a scenario in which the random field of interest  $\xi$  is an embedding of two random fields  $\xi_1$  and  $\xi_2$  with  $\xi_s \stackrel{\text{iid}}{\sim} g_s$  for  $s = 1, 2$ . The following normal densities  $N(0, 1)$ ,  $N(0.5, 1)$ ,  $N(1.5, 1.5)$ ,  $N(2.5, 1.5)$ ,  $N(3.5, 2)$ ,  $N(4.75, 2)$ ,  $N(6, 2.25)$ ,  $N(7, 2.25)$  are taken as the eight component densities  $\varphi_1(\xi), \dots, \varphi_8(\xi)$ . The same densities are used for both the training and test images, respectively. We have two class conditional pdfs:  $g_1(\xi) = \sum_{j=1}^6 \pi_j^1 \varphi_j(\xi)$  with  $\pi_j^1 = 1/6$  for  $j = 1, \dots, 6$  and  $g_2(\xi) = \sum_{j=4}^8 \pi_j^2 \varphi_j(\xi)$  with  $\pi_j^2 = 1/5$  for  $j = 4, \dots, 8$ . Figures 2(a) and (b) show the realizations of  $\xi_1$  and  $\xi_2$ , respectively. This setting might seem extreme, but it is useful to investigate the effects of highly overlapping densities with having many mixing coefficients on the boundary of the parametric space.

Let  $b$  be a binary (0, 1) Markov random field used in modeling the presence of different classes of objects. Figure 2(c) shows the realization of  $b$ , generated using a Gibbs sampler (Geman and Geman 1984), with an initial Bernoulli field (with  $p = 0.56$ ) and a 24-pixel square neighborhood. The random field of interest  $\xi$ , shown in Figure 2(d), is an embedding

of the two random fields  $\xi_1$  and  $\xi_2$  such that  $\xi_s \stackrel{\text{iid}}{\sim} g_s$  via the piecewise stationary random field  $\xi = I_{\{b(s)=0\}}\xi_1 + I_{\{b(s)=1\}}\xi_2$ . Thus, the random field  $\xi$  in Figure 2(d) is the union of  $r = 2$  disjoint regions and each region is composed of some disjoint and connected subregions. Note that  $\xi(s)$  is identically distributed as  $\xi_1(s)$  (or  $\xi_2(s)$ ) for  $s$  in the “white”



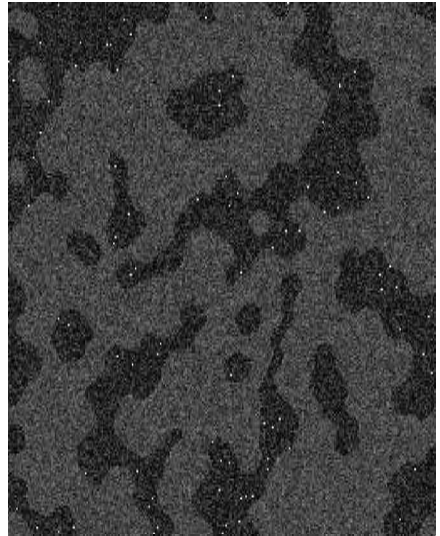
(a)



(b)



(c)



(d)

Figure 2. Realizations of stationary random fields: (a) Stationary random field  $\xi_1$ , (b) stationary random field  $\xi_2$ , (c) binary random field  $b$  used to embed  $\xi_1$  and  $\xi_2$  into  $\xi$  and (d) a realization of a piecewise stationary random field  $\xi$ .

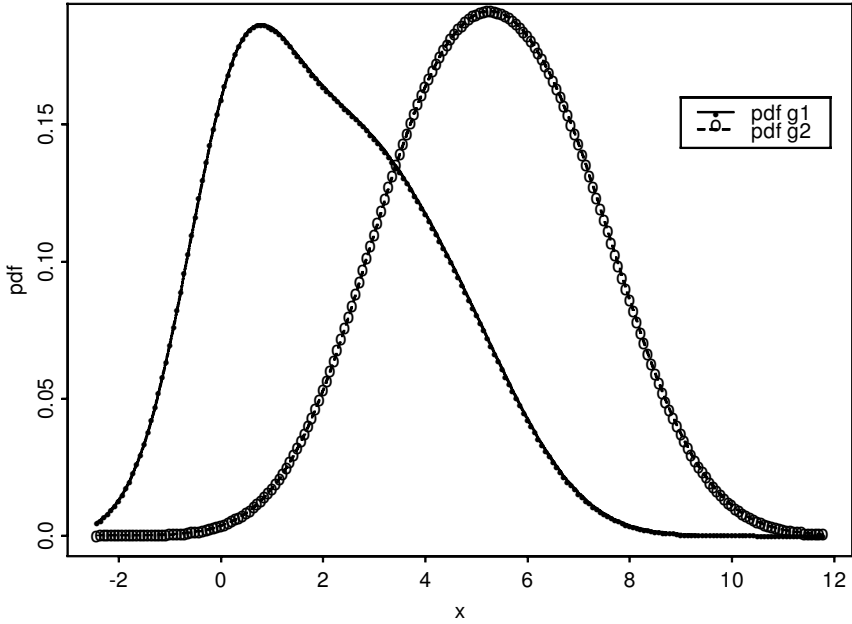


Figure 3. Class conditional pdfs  $g_1$  and  $g_2$ .

(or “black”) region associated with  $I_{\{b(s)=0\}}$  (or  $I_{\{b(s)=1\}}$ ). Let Image S1 be the Figure 2(d) which is a realization of the piecewise stationary random field embedded by the binary random field  $b$  in Figure 2(c). The class conditional pdfs  $g_1$  and  $g_2$  corresponding to  $\xi_1$  and  $\xi_2$ , respectively, are shown in Figure 3. Similarly, we generated Images S2 and S3 (not shown here) with different realizations of the binary field  $b$ .

The field  $\xi$  meets the criteria for applying the BSDE approach in detecting the targets. The scan density  $g_s$  associated with the scan window  $N_s(\delta)$  is estimated and the pixel  $s$  is then classified according to the rule:

$$s \in \text{class 1} \quad \text{if and only if} \quad \frac{d(\widehat{g}_s, g_1)}{d(\widehat{g}_s, g_2)} \leq T, \tag{6.1}$$

where  $d(\cdot, \cdot)$  is any distance and  $T$  is a prespecified threshold. In the present context, the integrated square error is used to measure the distance. That is,  $d(\widehat{g}_s, g) = \int_{-\infty}^{\infty} [\widehat{g}_s(\mathbf{y}) - g(\mathbf{y})]^2 dy$ . See Geman (1990) for other types of distance measures such as the relative entropy that one could use. We first generated the data Figure 2(c) to create Figure 2(d). The goal is to obtain Figure 2(c) from Figure 2(d) since in practice, one has only the knowledge of Figure 2(d) but not Figure 2(c). The classification of each pixel to its corresponding class via the rule (6.1) creates a figure similar to that of Figure 2(c). However, the resulting figure obtained from Figure 2(d) may not appear like Figure 2(c) due to noise and classification errors.

We consider two key measures for comparison of the two methods: (1) the average number of parameter updates per scan window required to reach the desired accuracy of  $\varepsilon$  and (2) the corresponding average CPU time per scan window relative to that of the

Table 1. Comparisons at Different Tolerance Levels of  $\varepsilon$  for a Fixed Scan Window Size of  $\delta = 2$ 

$\varepsilon$	<i>Simulation (Image S1)</i> <i>Improvement</i>		<i>Experiment (Image 1)</i> <i>Improvement</i>	
	<i>Speed</i>	<i># Updates</i>	<i>Speed</i>	<i># Updates</i>
0.1	7.0	4.8	3.0	3.7
0.01	13.3	7.1	3.7	5.2
0.005	16.2	8.4	4.3	5.7

conventional EM. The results are summarized in Table 1. The *Improvement* column gives the relative improvement of the adaptive SSD estimation method over the traditional BSDE approach with respect to the number of parameter updates and the inverse of the CPU time of the adaptive method. It is clear from Table 1 that the adaptive method always outperforms the traditional one with 4- to 8-fold improvement in terms of parameter updates and 7- to 16-fold improvement in terms of the speed. Note that the improvement is quite significant especially at  $\varepsilon = 0.005$ . Table 2 comparisons were done at  $\varepsilon = 0.01$  since at  $\varepsilon = 0.005$ , the traditional BSDE method was extremely slow, especially for larger window sizes, to obtain the results in a reasonable amount of time. Table 2 demonstrates the effect of increasing the window size  $\delta$  for a fixed  $\varepsilon$ . Once again, the magnitude of improvement obtained via the adaptive method depends on the size of the scan window and shows a 26-fold improvement for  $\delta = 4$ . In Table 3, we fix both  $\delta$  and  $\varepsilon$  at specified values and then assess the performance of the methods on various images.

It is not surprising to see that the choice of the method becomes clear as the window size increases or as the tolerance parameter decreases. This is expected since in the first case, the benefits of the dynamic adaptation rule (Section 4.1) become apparent with an increasing size of the overlapping region; whereas in the latter case, the higher the required accuracy, the slower the traditional BSD estimator is to progress towards the solution. The plots in Figure 4 confirm the results presented in the tables.

## 7. APPLICATION TO AUTOMATIC TARGET RECOGNITION

One of the goals of an ATR experiment is to detect and identify ROI in images. Given the image of a target region, spatial scan analysis based approach uses estimates obtained

Table 2. Comparisons at a Fixed Tolerance Level of  $\varepsilon$  and for Varied Scan Window Sizes

$\delta$	<i>Simulation (Image S1)</i> $\varepsilon = 0.01$ <i>Improvement</i>		<i>Experiment (Image 1)</i> $\varepsilon = 0.005$ <i>Improvement</i>	
	<i>Speed</i>	<i># Updates</i>	<i>Speed</i>	<i># Updates</i>
1	9.3	5.8	3.6	5.2
2	13.3	7.1	4.3	5.7
3	17.4	9.3	4.4	6.0
4	26.4	13.2	5.0	6.2



Table 3. Comparisons at a Fixed Scan Window Size of  $\delta = 2$  for Different Images

<i>Simulation (<math>\epsilon = 0.01</math>)</i>			<i>Experiment (<math>\epsilon = 0.005</math>)</i>		
<i>Image identity</i>	<i>Improvement</i>		<i>Image identity</i>	<i>Improvement</i>	
	<i>Speed</i>	<i># Updates</i>		<i>Speed</i>	<i># Updates</i>
S1	13.3	7.1	1	4.3	5.7
S2	13.6	7.3	2	5.7	7.5
S3	13.5	7.3	3	4.1	5.1
—	—	—	4	3.8	4.2
—	—	—	5	3.6	4.7
—	—	—	6	4.4	5.5

via the moving scan window to estimate the image characteristics in local regions. In turn, these local estimates are compared across the entire image and if the local estimates are nearly identical, then the region is considered as homogeneous and hence contains no ROI. However, if the local estimate is significantly different from the majority, it is labeled as a potential region of interest for further study.

We now investigate the new approach via the ATR experiment in gray-scale images; in particular, the detection and identification of man-made ROI in UAV imagery. A set of six infrared images, shown in Figure 5, extracted from a videotape of a UAV test flight over the Naval Strike Warfare Center, Fallon, Nevada in the summer of 1995 (Solka et al. 1998) are used in our study. In here, the task of an analyst is to determine the location of possible

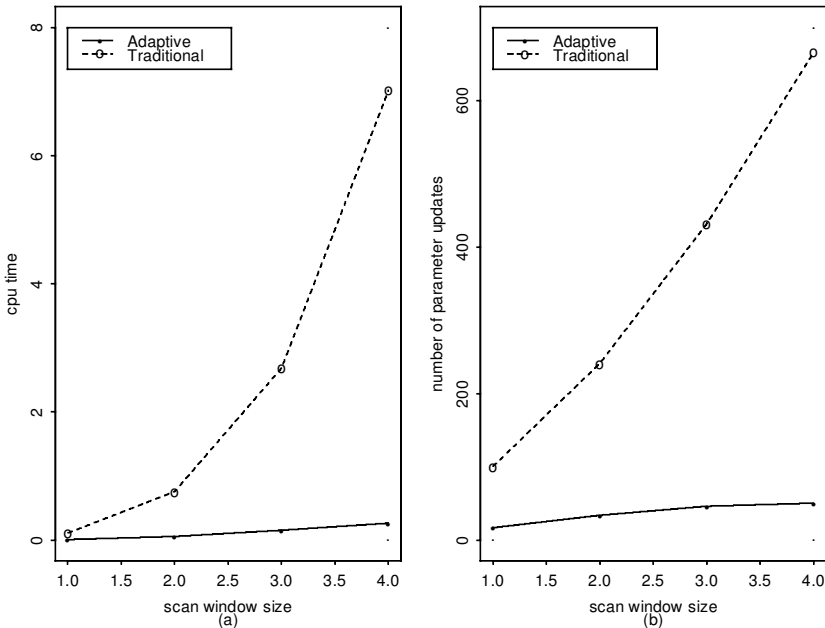


Figure 4. Performance curves of the scan processes at the tolerance level of  $\epsilon = 0.01$  for varied scan window sizes: (a) average CPU time per scan window and (b) average number of parameter updates per scan window.

targets. We build a training model (see Figure 6) as a four-component mixture model based on Image 1, a training image.

In assessing the methods, we again consider three cases (1) varying  $\delta$  for a fixed  $\varepsilon$ , (2) varying  $\varepsilon$  for a fixed  $\delta$ , and (3) varying images for a fixed  $\delta$  and  $\varepsilon$ . From Tables 1–3 and Figure 7 it is clear that the adaptive SSD estimation method outperforms the traditional one. The findings are similar to the simulation study, although the improvements are not as dramatic.



Image 1

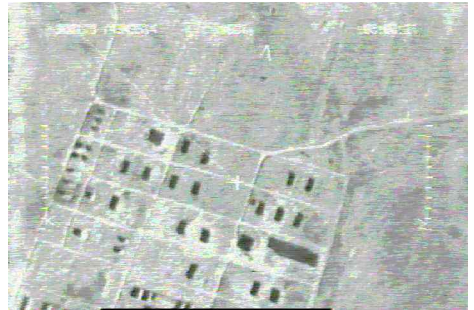


Image 2



Image 3

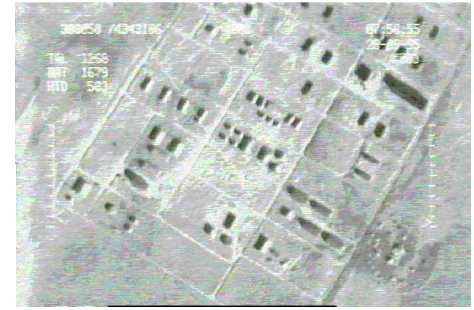


Image 4



Image 5

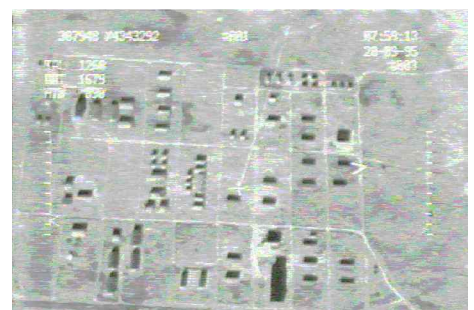


Image 6

Figure 5. Zoomed out electro optical images of the army compound.

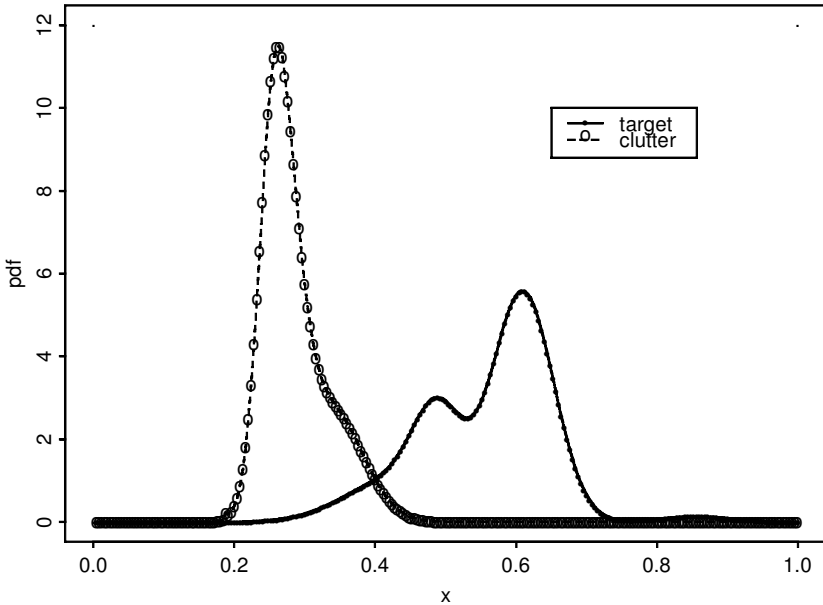


Figure 6. Training models from Image 1.

However, the ATR experiment is conducted in real-time and hence an improvement of five-fold is still substantial.

## 8. DISCUSSION

This article proposes efficient SSD estimation methods by integrating the borrowed strength technique into the alternative complete data framework. The new methods combine the statistical basis of the BSDE (Section 2) procedure with the stability and improved convergence rate of the alternative EM methods (Section 3). Furthermore, we extended the aforementioned approaches via a dynamic adaptation rule for the mixing coefficients (Section 4) to eliminate the need to find the posterior probability of membership of the component densities afresh in each subregion. The resulting adaptive SSD estimation methods (a) provide fast processing; (b) naturally adapt to changes in local image variations (often the case in missile guidance systems) and background noise as well as the uncertainty of target properties; and (c) improve the speed of the ATR methods by several fold, with gains that increase as the size of the scan window increases or as the level of tolerance parameter for convergence decreases. The adaptive method based on the Rotated EM algorithm is not only simple and flexible but also showed a seven to 26-fold improvement in simulation studies (Section 6) and a three to five-fold improvement in the ATR experiment (Section 7); with greater improvements for larger window sizes. Improvements were the greatest in the case of highly overlapping densities. It is possible to apply our adaptive SSD esti-

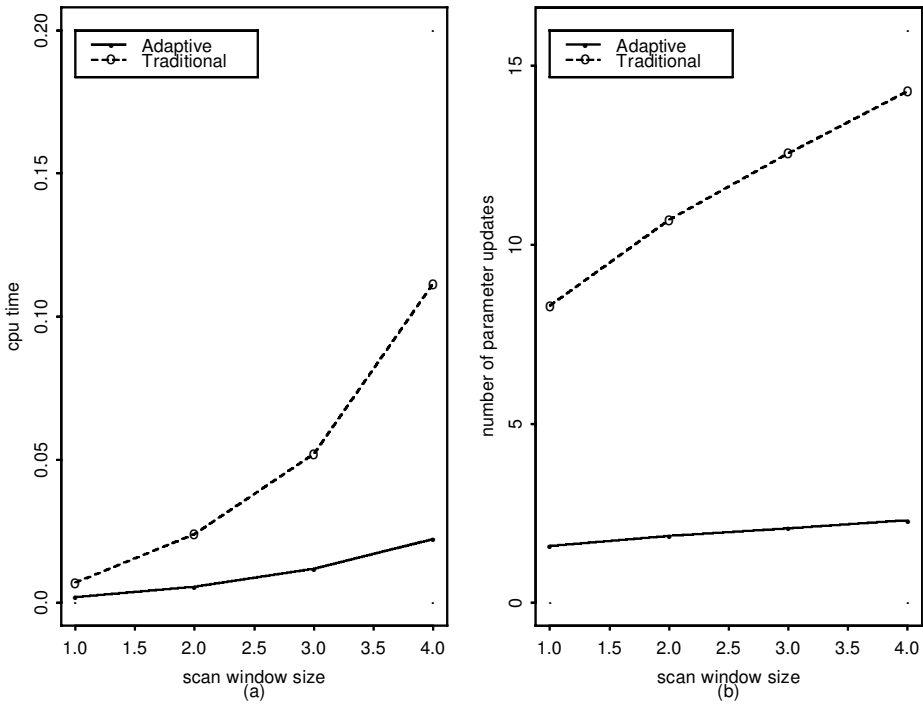


Figure 7. Performance curves of scan processes at  $\varepsilon = 0.01$ : (a) average CPU time per scan window and (b) average number of parameter updates per scan window.

mation methods in contexts where great speed is required such as in processing real-time and interactive images. We believe that the methods presented in this article represent a substantial advance for a general larger (possibly complex) class of image problems and spatial analysis.

We would not expect a substantial improvement over the traditional BSDE approach if the densities were well separated, overlapping region between the adjacent scan windows were minimal, or most of the mixing coefficients were positive. Due to the simplicity of the BSDE via the Rotated EM, one might prefer it; however, when most of the weight parameters are zeros, one might combine the Rotated EM with the zero-elimination scheme (by sequentially eliminating those support points from the algorithm that have mass zero) proposed by Pilla and Lindsay (2001).

The dynamic adaptation rule developed for the mixture problem in Section 4 can be generalized to any problem that can be cast in the missing data framework. The amount of missing data is directly related to the rate of convergence of an EM-based algorithm. Hence it is natural to avoid replacing the entire missing data with their conditional expectations when indeed there is only a fraction of which is missing in the overlapping region. We believe that this enables one to adaptively manage the missing data in the overlapping region and further improve the computational efficiency.

An alternative approach to the adaptive SSD estimation would replace the AKM algo-

rithm with the nonparametric maximum likelihood (NPML) estimation, where we maximize the likelihood over all the mixing distributions. As pointed out by Pilla and Lindsay (2001), one can start with a Rotated EM on a fine grid of  $\theta$  parameters (act as the support set of the mixing distribution) and in turn use the resulting solution as a starting point for the continuous support EM (estimates the  $\theta$  parameters simultaneously with the  $\pi$  parameters). This eliminates the problem of using incorrect number of support points or finding suboptimal solutions. See Pilla and Lindsay for details on finding the NPML estimator via a fixed support mixture algorithm.

## ACKNOWLEDGMENTS

Pilla's research was supported in part by the NSF grant DMS 02-39053 and Probability and Statistics Program, Office of Naval Research (ONR) grant N00014-02-1-0316. Priebe's research was partially supported by the ONR Grant N00014-01-1-0011. The authors thank David Marchette and Jeffrey Solka of the Naval Surface Warfare Center for their generosity in providing the unmanned aerial vehicle test flight data analyzed in Section 7. We also thank the editor David Scott, associate editor, and the referees for their constructive comments. Authors are grateful to David Marchette for many stimulating discussions.

[Received March 2001. Revised January 2002.]

## REFERENCES

- Bhanu, B., Dudgeon, D. E., Zelnio, E. G., Rosenfeld, A., Casasent, D., and Reed, I. S. (1997), "Introduction to the Special Issue on Automatic Target Detection and Recognition," *IEEE Transactions on Image Processing*, 6, 1–7.
- Böhning, D., and Lindsay, B. G. (1988), "Monotonicity of Quadratic-Approximation Algorithms," *Annals of the Institute of Statistical Mathematics*, 40, 641–663.
- Chen, J., and Glaz, J. (1997), "Two-Dimensional Discrete Scan Statistics," *Statistics and Probability Letters*, 31, 59–68.
- Cox, D. R., and Reid, J. (1987), "Parameter Orthogonality and Approximate Conditional Inference" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 49, 1–39.
- Cressie, N. A. C. (1993), *Statistics for Spatial Data*, (revised ed.), New York: Wiley.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, 39, 1–22.
- Forbes, F. and Raftery, A. E. (1999), "Bayesian Morphology: Fast Unsupervised Bayesian Image Analysis," *Journal of the American Statistical Association*, 94, 555–568.
- Geman, D. (1990), *Random Fields and Inverse Problems in Imaging*, Lecture Notes in Mathematics, Berlin: Springer Verlag.
- Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.
- IEEE Transactions on Image Processing* (1997), 6, Special Issue on Automatic Target Recognition.
- Lange, K. (1995), "A Gradient Algorithm Locally Equivalent to the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, 57, 425–437.
- Lindsay, B. G. (1995), *Mixture Models: Theory, Geometry and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, Vol 5., Haywood, CA: Institute of Mathematical Statistics.

- Marchette, D. J., Priebe, C. E., Rogers, G. W., and Solka, J. L. (1996), "Filtered Kernel Density Estimation," *Computational Statistics*, 11, 95–112.
- McLachlan, G. J., and Krishnan, T. (1997), *The EM Algorithm and Extensions*, New York: Wiley.
- McLachlan, G. J., and Peel, D. (2001), *Finite Mixture Models*, New York: Wiley.
- Pilla, R. S. (1997), "Improving the Rate of Convergence of the EM in High Dimensional Finite Mixtures," unpublished Ph.D. dissertation, Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania.
- Pilla, R. S., and Lindsay, B. G. (1996), "Fast EM Methods in High Dimensional Finite Mixtures," in *Proceedings of the Statistical Computing Section*, Alexandria, VA: American Statistical Association, pp. 166–171.
- (2001), "Alternative EM Methods for Nonparametric Finite Mixture Models," *Biometrika*, 88, 535–550.
- Popat, K., and Picard, R. W. (1997), "Cluster-based Probability Model and its Application to Image and Texture Processing," *IEEE Transactions on Image Processing*, 6, 268–284.
- Priebe, C. E. (1996), "Nonhomogeneity Analysis Using Borrowed Strength," *Journal of the American Statistical Association*, 91, 1497–1503.
- Priebe, C. E., and Chen, D. (2001), "Spatial Scan Density Estimates," *Technometrics*, 43, 73–83.
- Priebe, C. E., and Marchette, D. J. (2000), "Alternating Kernel and Mixture Density Estimation," *Computational Statistics and Data Analysis*, 35, 43–65.
- Priebe, C. E., Marchette, D. J., and Rogers, G. W. (1997a), "Segmentation of Random Fields via Borrowed Strength Density Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 494–499.
- Priebe, C. E., Solka, J. L., and Tao, P. (1997b), "Spatial Scan Analysis of Unmanned Aerial Vehicle Imagery," *Proceedings of the Computing Section*, Alexandria, VA: American Statistical Association, pp. 92–97.
- Solka, J. L., Marchette, D. J., Wallet, B. C., Irwin, V. L., and Rogers, G. W. (1998), "Identification of Man-made Regions in Unmanned Aerial Vehicle Imagery and Videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20, 852–857.
- Tao, P. (2000), "The Generalized Borrowed Strength Method and the Application to Image Recognition," unpublished Ph.D. dissertation, Department of Mathematical Sciences, Johns Hopkins University, Baltimore, Maryland.
- Titterton, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.