http://www.jstor.org

Spatial Scan Density Estimates

Carey E. PRIEBE AND Dalei CHEN

Department of Mathematical Sciences
The Johns Hopkins University
Baltimore, MD 21218-2682
(*cep@jhu.edu*)
(*dalei.chen@bms.com*)

The detection and identification of regions of interest in spatial or temporal data is a common concern in automatic target recognition. One approach to region-of-interest identification involves the use of spatial scan statistics. A difficulty arises due to competing concerns: Small scan windows are required for potentially small targets, but larger scan windows are necessary to improve the accuracy of the detector. When the scan statistics are mixture-model density estimates, a borrowed strength profile likelihood approach is shown to be superior to conventional likelihood estimators. We investigate these spatial scan density estimates on example imagery from an unmanned aerial vehicle.

KEY WORDS: Automatic target recognition; Borrowed strength; Density estimation; Mixture models; Nonhomogeneity detection; Profile likelihood; Region-of-interest identification; Spatial scan analysis; Texture features; Unmanned aerial vehicle imagery.

Automatic target recognition (ATR) is a problem of great importance attracting significant research activity; see, for instance, the special issue on ATR of *IEEE Transactions on Image Processing* (1997). One aspect of ATR is the automated detection and identification of regions of interest (ROI) in images. Given the image of a target area, our approach is that of spatial scan analysis, or "moving window estimates," wherein image characteristics in local regions are estimated and these estimates are compared across the entire image. If all the local estimates are nearly identical, then the area is likely homogeneous, indicating that it contains no ROI. A region for which the local estimate is significantly different from the majority "background" should be labeled as a potential region of interest for further study.

Let an image be represented as real-valued observations $\xi(\mathbf{x})$ made at spatial locations $\mathbf{x} \in R^0 \subset \mathbb{R}^d$, where the observations represent pixel intensity. Cressie (1993) described the *scan process* for given $\delta > 0$ as $f(\xi(\mathbf{y}) : \mathbf{y} \in R(\mathbf{x}; \delta))$ for all $\mathbf{x}$ such that $R(\mathbf{x}; \delta) \subset R^0$. The $R(\mathbf{x}; \delta)$ are *scan regions* around the spatial locations $\mathbf{x}$. For a discrete image defined at pixel locations on a lattice, we let I and n^i denote the number of scan regions and the number of observations in the ith region R^i respectively. The *locality statistic* $f(\cdot)$ is a function of the observations in the local neighborhood around $\mathbf{x}$. One choice for $R(\mathbf{x}; \delta)$ is the ball of radius δ centered at $\mathbf{x}$.

Henceforth, we use ξ instead of $\xi(\mathbf{x})$ for simplicity. We consider the local regional image characteristics of interest to be the parameters in a finite mixture estimate of the local marginal probability density function. That is, the probability density function α^i has the functional form

$$\alpha^i(\xi) = \alpha(\xi; \theta^i) = \sum_{t=1}^{m_i} \lambda_t^i C(\xi; \psi_t^i), \qquad (1)$$

where each λ_t^i is nonnegative, $\sum_{t=1}^{m_i} \lambda_t^i = 1$, and each $C(\xi; \psi_t^i)$ is itself a density function parameterized by $\psi_t^i \in \Psi_t^i \subset \mathbb{R}^{k_t}$. We denote $\theta^i = (\psi^i, \lambda^i) = ((\psi_1^i, \ldots, \psi_{m_i}^i), (\lambda_1^i, \ldots, \lambda_{m_i}^i))'$

and set

$$\Theta^i = \{(\psi^i, \lambda^i)' : 0 \le \lambda_t^i \le 1, \sum_{t=1}^{m_i} \lambda_t^i = 1,$$
$$\text{and } \psi_t^i \in \Psi_t^i \text{ for all } t = 1, \ldots, m_i\}.$$

(Here and hereafter we take $t = 1, \ldots, m_i - 1$ in the mixing coefficients λ_t^i due to the constraint that the λ_t^i sum to unity.)

We use an estimate of α^i, $\hat{\alpha}^i = \alpha(\cdot; \hat{\theta}^i) = f(\xi(\mathbf{x}) : \mathbf{x} \in R^i)$, as the locality statistic. A fundamental assumption underlying our approach to the ROI problem is that the (marginal) probability density function of local image features can be used to distinguish between background and regions of interest.

Our interest is to determine if there are candidate ROI in the image or if the I regions are identical in terms of the image characteristics under consideration. In this framework, the null hypothesis in a traditional multiple-comparisons methodology is that all the marginal density functions α^i are the same, whereas the alternative hypothesis is the nonhomogeneity among those marginal densities.

The key to the test of nonhomogeneity is to find an accurate estimate of the maximum pairwise distance $D = \max_{i,j \in \{1, \ldots, I\}} d(\alpha^i, \alpha^j)$, where the (pseudo-)distance $d(\cdot, \cdot)$ is defined on the space of probability densities under consideration. Large values of D indicate nonhomogeneity of the image, thus giving evidence of the existence of an ROI.

To test for nonhomogeneity, we consider the scan statistic $T = \max_{i,j \in \{1, \ldots, I\}} d(\hat{\alpha}^i, \hat{\alpha}^j)$. The power of the test based on this statistic is related to the success of T as an estimate of D. Regardless of the choice for $d(\cdot, \cdot)$, the variance of the estimates $\hat{\alpha}^i$ will affect the variance of the statistic T, which in turn may affect its power to detect large distances between two densities. Because the densities α^i are finite mixture models, the variance of T is a function of the variance of the parameter

estimates $\hat{\theta}^i$. Thus, better estimates of θ^i yield a more accurate scan statistic T that should have higher power for the test of nonhomogeneity under the alternative.

Different choices are available for $d(\cdot, \cdot)$. In Section 2, we consider tests for which $d(\cdot)$ is based on the squared error (SE) of the mixing coefficients; given the similarity assumption $\psi^i = \psi^j$ for all i, j discussed in Section 2, we present results based on $SE(\lambda^i, \lambda^j) = (\lambda^i - \lambda^j)'(\lambda^i - \lambda^j)$. Section 2.3 relates these results to tests based on integrated squared error (ISE):

$$ISE(\theta^i, \theta^j) = \int (\alpha(\xi; \hat{\theta}^i - \alpha(\xi; \theta^j)))^2 d\xi.$$

The choice of $d(\cdot, \cdot)$ is an important issue in the test of nonhomogeneity. If this choice does not capture important differences in the local image characteristics, high power of the test of nonhomogeneity will not necessarily translate into better detection of ROI.

Given an appropriate measure of distance $d(\cdot, \cdot)$, the efficiency of detection based on spatial scan analysis is highly influenced by the variance and bias of the estimates. To detect anticipated ROI of unknown size, the size of the scan regions needs to be relatively small. Thus, the number of observations n^i in each scan region, on which the local estimates are based, is limited. This in turn suggests that the local estimates may have unacceptably large bias and variance, which reduces the probability of the scan statistic T being close to D, thus adversely affecting the power of the test for nonhomogeneity. Dependency further exacerbates this problem; because imagery exhibits (at least local) positive correlation, the *effective* local sample size, in terms of equivalent number of independent observations (Cressie 1993), is even smaller than n^i. (A second dependency, due to the overlap in adjacent scan regions, complicates the multiple-comparison problem.)

To address the conundrum raised by the competing requirements for small scan regions (yielding small sample sizes) and large sample sizes (for better local estimates), we consider the borrowed strength methodology (Priebe 1996). This entails combining all the observations in R^0 into one sample and estimating an unconditional probability density function. The estimates of the regional characteristics are obtained through a profile likelihood technique. An assumption for using the borrowed strength methodology is that the underlying marginal densities of all the regions can be modeled with finite mixture models. In addition, the difference between the mixture models for the different classes is encompassed entirely in the mixing coefficients. These assumptions are practical because mixture models can approximate (nearly) any probability density function (e.g., Young and Coraluppi 1970) and even low-order normal mixtures span a rich class of densities (Marron and Wand 1992).

We show that, by reducing the mean squared error (MSE) of the local parameter estimates $\hat{\psi}^i$ and $\hat{\lambda}^i$, the borrowed strength methodology is superior to the local likelihood (LL) methodology, which uses only observations in one scan region for the corresponding local estimate. Furthermore, we compare the borrowed strength estimates with those obtained via the seemingly appropriate joint likelihood (JL) approach and exhibit the superiority of the former.

The LL, JL, and borrowed strength profile likelihood (BSPL) functions are given in Section 1, together with the steps of the EM algorithm for obtaining the maximum likelihood estimate (MLE) based on each. Section 2 discusses the properties of these three estimators and, in particular, compares the JL with borrowed strength in terms of the asymptotic bias of the JL estimates $\hat{\psi}_{JL}^0$ and the difference between the asymptotic MSE of the JL estimate $\hat{\lambda}_{JL}^i$ and the BSPL estimate $\hat{\lambda}_{PL}^i$. The intuition for the superiority of the borrowed strength estimates comes from the nonorthogonality between parameters (see Cox and Reid 1987). Here we obtain a proof of the asymptotic biasedness of the joint likelihood estimate, in the spirit of Neyman and Scott (1948), by analyzing the EM steps. In Section 3 we consider the application of our methodology to the ATR ROI identification problem. We conclude in Section 4 with a discussion of the results, their implications, and their applicability.

1. FINITE MIXTURE ESTIMATES IN SPATIAL SCAN ANALYSIS

For the spatial scan process under consideration, we assume that the observations in region R^i, $\{\xi_j^i = \xi(x_j^i), j = 1, \ldots, n^i\}$, have the identical underlying density α^i. For practical purposes, it is straightforward to incorporate dependence structures such as m-step dependence and simple mixing into the estimation frameworks considered in this section (Guyon 1995). The dependency translates into a smaller effective sample size but does not otherwise affect the asymptotic behavior of the estimators. The likelihood functions become pseudolikelihoods. For simplicity and concreteness, we assume in what follows that the observations ξ_j^i are mutually independent.

We consider three competing estimates for α^i that are relevant to spatial scan applications. These estimates are MLE's obtained from the LL, JL, and BSPL functions, respectively.

1.1 Local Likelihood

Consideration of only the local sample, the observations in region R^i, yields the simplest of the three estimates. Based on the observations ξ_j^i and the marginal density α^i given in (1), the regional LL function is

$$L^i(\theta^i) = \prod_{j=1}^{n^i} \alpha^i(\xi_j^i) = \prod_{j=1}^{n^i} \alpha(\xi_j^i; \theta^i) \tag{2}$$

and the MLE therefrom, hereafter called the LL estimate, is denoted as $\hat{\theta}_{LL}^i = (\hat{\psi}_{LL}^i, \hat{\lambda}_{LL}^i)$. One such estimate is required for each scan region R^i.

The EM algorithm (Dempster, Laird, and Rubin 1977; Redner and Walker 1984; McLachlan and Krishnan 1997) is applied to obtain the LL estimate $\hat{\theta}_{LL}^i$. For normal mixtures [i.e., in (1), $C(\cdot)$ is a normal density $\phi(\cdot)$ with $\psi_t^i = (\mu_t^i, \nu_t^i)'$, where μ_t^i and ν_t^i are the mean and variance, respectively, for the tth mixture component in the ith scan region], the EM

algorithm gives the iterative parameter update equations

$$\hat{\lambda}_t^{i(k+1)} = \frac{1}{n^i} \sum_{j=1}^{n^i} \hat{\rho}_{jt}^{i(k)},$$

$$\hat{\mu}_t^{i(k+1)} = \frac{\sum_{j=1}^{n^i} \xi_j^i \hat{\rho}_{jt}^{i(k)}}{n^i \hat{\lambda}_t^{i(k)}} \quad \text{and}$$

$$\hat{\nu}_t^{i(k+1)} = \frac{\sum_{j=1}^{n^i} \hat{\rho}_{jt}^{i(k)} \left(\xi_j^i - \hat{\mu}_t^{i(k+1)} \right)^2}{n^i \hat{\lambda}_t^{i(k)}},$$

where

$$\hat{\rho}_{jt}^{i(k+1)} = \frac{\hat{\lambda}_t^{i(k)} \phi\left(\xi_j^i; \hat{\mu}_t^{i(k)}, \hat{\nu}_t^{i(k)} \right)}{\sum_{l=1}^{m_i} \hat{\lambda}_l^{i(k)} \phi(\xi_j^i; \hat{\mu}_l^{i(k)}, \hat{\nu}_l^{i(k)})},$$

for $i = 1, \ldots, I$, $j = 1, \ldots, n^i$, and $t = 1, \ldots, m_i$.

1.2 Joint Likelihood

To apply the JL methodology, one must assume common characteristics across the scan regions. This commonality is achieved by letting all the mixtures be of the same order ($m_i = m$ for $i = 1, \ldots, I$) and designating $\boldsymbol{\psi}^i$ to be *structural* parameters (Neyman and Scott 1948) with $\boldsymbol{\psi}^i = \boldsymbol{\psi}^0$ for all i.

Thus,

$$\alpha^i(\xi) = \alpha(\xi; \boldsymbol{\theta}^i) = \alpha(\xi; \boldsymbol{\psi}^0, \boldsymbol{\lambda}^i) = \sum_{t=1}^{m} \lambda_t^i C(\xi; \boldsymbol{\psi}_t^0)$$

so that the difference between the densities is encompassed entirely in the *incidental* parameters (Neyman and Scott 1948)—the mixing coefficients $\boldsymbol{\lambda}^i = (\lambda_1^i, \ldots, \lambda_{m-1}^i)'$. For example, for normal mixtures, Equation (1) becomes

$$\alpha^i(\xi) = \alpha(\xi, \boldsymbol{\psi}^0, \boldsymbol{\lambda}^i)$$

$$= \alpha\left(\xi; (\mu_1^0, \nu_1^0, \ldots, \mu_m^0, \nu_m^0), (\lambda_1^i, \ldots, \lambda_{m-1}^i)\right)$$

$$= \sum_{t=1}^{m} \lambda_t^i \phi(\xi; \mu_t^0, \nu_t^0), \tag{3}$$

where $\boldsymbol{\psi}^0 = (\mu_1^0, \nu_1^0, \ldots, \mu_m^0, \nu_m^0)'$ is the common structural parameter to all of the densities α^i.

Because the structural parameter vector $\boldsymbol{\psi}^0$ is common to all the regions, it is natural to consider the JL estimate $\hat{\boldsymbol{\theta}}_{\text{JL}}^i = (\hat{\boldsymbol{\psi}}_{\text{JL}}^0, \hat{\boldsymbol{\lambda}}_{\text{LL}}^i)$ obtained by maximizing the JL function

$$L^J(\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^I) = \prod_{i=1}^{I} \prod_{j=1}^{n^i} \alpha(\xi_j^i; \boldsymbol{\theta}^i) = \prod_{i=1}^{I} \prod_{j=1}^{n^i} \alpha(\xi_j^i; \boldsymbol{\psi}^0, \boldsymbol{\lambda}^i).$$

$$\tag{4}$$

The EM algorithm for the JL estimate for m-component normal mixtures yields

$$\check{\lambda}_t^{i(k+1)} = \frac{1}{n^i} \sum_{j=1}^{n^i} \check{\rho}_{jt}^{i(k)},$$

$$\check{\mu}_t^{(k+1)} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n^i} \xi_j^i \check{\rho}_{jt}^{i(k)}}{\sum_{i=1}^{I} n^i \check{\lambda}_t^{i(k)}},$$

and

$$\check{\nu}_t^{(k+1)} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n^i} \check{\rho}_{jt}^{i(k)} (\xi_j^i - \check{\mu}_t^{(k+1)})^2}{\sum_{i=1}^{I} n^i \check{\lambda}_t^{i(k)}},$$

where

$$\check{\rho}_{jt}^{i(k+1)} = \frac{\check{\lambda}_t^{i(k)} \phi(\xi_j^i; \check{\mu}_t^{(k)}, \check{\nu}_t^{(k)})}{\sum_{l=1}^{m} \check{\lambda}_l^{i(k)} \phi(\xi_j^i; \check{\mu}_l^{(k)}, \check{\nu}_l^{(k)})}$$

for $i = 1, \ldots, I$, $j = 1, \ldots, n^i$, and $t = 1, \ldots, m$.

Under the similarity assumption, the JL estimate is more appropriate than the LL estimate because it uses observations in all the regions to estimate the structural parameter $\boldsymbol{\psi}^0$. A major deficiency of JL is that it is appropriate only for nonoverlapping scan regions. When overlapping scan windows are used, which is advantageous because it eliminates the boundary issue inherent in an arbitrary choice of partition, the applicability of the JL procedure is limited. [Note that when the sizes of the scan windows are all the same and edge effects are disregarded, the log of Eq. (4) is proportional to the actual log JL.]

1.3 Borrowed Strength Profile Likelihood

Under the same assumptions considered in the previous section, we now present BSPL. If we consider that any of the $n^0 = \sum_i n^i$ observations are taken from the ith sample with probability p^i proportional to n^i for all $i = 1, \ldots, I$, $\sum_i p^i = 1$, then the unconditional [unconditional on the location $\mathbf{x}$ of $\xi(\mathbf{x})$] density function can be expressed as

$$\alpha^0(\xi) = \sum_{i=1}^{I} p^i \alpha^i(\xi) = \sum_{t=1}^{m} \lambda_t^0 \phi(\xi; \mu_t^0, \nu_t^0) \equiv \alpha(\xi; \boldsymbol{\psi}^0, \boldsymbol{\lambda}^0),$$

where $\boldsymbol{\lambda}^0 = \sum_i p^i \boldsymbol{\lambda}^i$.

The BSPL estimate is obtained by combining the likelihood estimate $\hat{\boldsymbol{\psi}}_{\text{PL}}^0$ obtained by maximizing

$$L^0(\boldsymbol{\psi}^0, \boldsymbol{\lambda}^0) = \prod_{j=1}^{n^0} \alpha(\xi_j; \boldsymbol{\theta}^0) = \prod_{j=1}^{n^0} \alpha(\xi_j; \boldsymbol{\psi}^0, \boldsymbol{\lambda}^0) \tag{5}$$

with regional profile likelihood (PL) estimates (McCullagh and Nelder 1989) $\hat{\boldsymbol{\lambda}}_{\text{PL}}^i$ obtained by maximizing the PL function

$$L^i(\boldsymbol{\lambda}^i | \hat{\boldsymbol{\psi}}_{\text{PL}}^0) = \prod_{j=1}^{n^i} \alpha(\xi_j^i; \boldsymbol{\lambda}^i | \hat{\boldsymbol{\psi}}_{\text{PL}}^0). \tag{6}$$

The first step in the BSPL demonstrates the idea of "borrowing strength." Under the similarity assumption, the structural parameter $\boldsymbol{\psi}^0$ is the same throughout the region. Thus, all the observations can be used to estimate $\boldsymbol{\psi}^0$.

In this case, the overall sample can be represented as $\bigcup_{i=1}^{I} \{\xi_j^i : j = 1, \ldots, n^i\} = \{\xi_n : n = 1, \ldots, n^0\}$. We use the EM algorithm twice first to find $\hat{\boldsymbol{\theta}}^0 = (\hat{\boldsymbol{\psi}}_{\text{PL}}^0, \hat{\boldsymbol{\lambda}}^0)$ and then to find $\hat{\boldsymbol{\lambda}}_{\text{PL}}^i$ given $\hat{\boldsymbol{\psi}}_{\text{PL}}^0$.

For normal mixtures the EM steps for $\hat{\boldsymbol{\psi}}_{\text{PL}}^0$ are essentially the same as the LL steps, except for using the overall sample

$$\hat{\lambda}_t^{0(k+1)} = \frac{1}{n^0} \sum_{i=1}^{I} \sum_{j=1}^{n^i} \hat{\rho}_{jt}^{i(k)} = \frac{1}{n^0} \sum_{n=1}^{n^0} \hat{\rho}_{nt}^{(k)},$$

$$\hat{\mu}_t^{0(k+1)} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n^i} \xi_j^i \hat{\rho}_{jt}^{i(k)}}{n^0 \hat{\lambda}_t^{0(k)}} = \frac{\sum_{n=1}^{n^0} \xi_n \hat{\rho}_{nt}^{(k)}}{n^0 \hat{\lambda}_t^{0(k)}},$$

and

$$\hat{\nu}_t^{0\,(k+1)} = \frac{1}{n^0 \hat{\lambda}_t^{0\,(k)}} \sum_{i=1}^{I} \sum_{j=1}^{n^i} \hat{\rho}_{jt}^{i\,(k)} (\xi_j^i - \hat{\mu}_t^{0\,(k+1)})^2$$

$$= \frac{1}{n^0 \hat{\lambda}_t^{0\,(k)}} \sum_{n=1}^{n^0} \hat{\rho}_{nt}^{(k)} (\xi_n - \hat{\mu}_t^{0\,(k+1)})^2,$$

where

$$\hat{\rho}_{jt}^{i\,(k+1)} = \frac{\hat{\lambda}_t^{0\,(k)} \phi(\xi_j^i; \hat{\mu}_t^{0\,(k)}, \hat{\nu}_t^{0\,(k)})}{\sum_{l=1}^{m} \hat{\lambda}_l^{0\,(k)} \phi(\xi_j^i; \hat{\mu}_l^{0\,(k)}, \hat{\nu}_l^{0\,(k)})} = \hat{\rho}_{nt}^{(k+1)},$$

for $i = 1, \ldots, I$, $j = 1, \ldots, n^i$, and $t = 1, \ldots, m$.

Once we obtain $\hat{\psi}_{PL}^0 = (\hat{\mu}_{1PL}^0, \hat{\nu}_{1PL}^0, \ldots, \hat{\mu}_{mPL}^0, \hat{\nu}_{mPL}^0)'$, the EM algorithm for $\hat{\lambda}_{PL}^i$ yields

$$\tilde{\lambda}_t^{i(k+1)} = \frac{1}{n^i} \sum_{j=1}^{n^i} \tilde{\rho}_{jt}^{i\,(k)},$$

where

$$\tilde{\rho}_{jt}^{i\,(k+1)} = \frac{\tilde{\lambda}_t^{i\,(k)} \phi(\xi_j^i; \tilde{\mu}_t^0, \tilde{\nu}_t^0)}{\sum_{l=1}^{m} \tilde{\lambda}_l^{i\,(k)} \phi(\xi_j^i; \tilde{\mu}_l^0, \tilde{\nu}_l^0)}$$

for $i = 1, \ldots, I$, $j = 1, \ldots, n^i$, and $t = 1, \ldots, m - 1$.

The BSPL methodology thus combines the structural parameter estimates based on all available observations with subsequent local profile estimates of the incidental parameters.

Intuitively, the BSPL estimates are superior to the JL estimates. Although the JL estimate of ψ^0 uses all the available observations as well, the incidental parameters λ^i are estimated at the same time. As suggested by Neyman and Scott (1948) and proved in Section 2, the JL estimate $\hat{\psi}_{JL}^0$ of the structural parameter ψ^0 is asymptotically biased for large I, the number of scan regions, and fixed n^i, the sample size. Cox and Reid (1987) indicated that the bias of the MLE of ψ^0 can adversely influence the estimate of λ^i if the two parameters are nonorthogonal. To resolve the problem of asymptotic biasedness, we introduce the unconditional marginal density α^0 with universal parameters (ψ^0, λ^0) to obtain an asymptotic unbiased estimate $\hat{\psi}_{PL}^0$, and this is then used in estimating local λ^i's, yielding a better estimate of λ^i. In Section 2, we compare the three estimation approaches in terms of the MSE of the various parameters.

2. COMPARISONS

Spatial scan analysis for ROI identification in ATR demands many small scan regions. Thus, the relevant asymptotic results are for $I \to \infty$ and n^i fixed for $i = 1, \ldots, I$. Let the MSE of an estimate of θ^i, say $\bar{\theta}^i$, be defined as $\text{MSE}(\bar{\theta}^i) = E(\text{se}(\bar{\theta}^i, \theta^i))$. In this section, the three competing likelihood estimators are investigated in terms of MSE as $I \to \infty$.

2.1 Comparison Between the BSPL and the LL Estimates

2.1.1 Local Likelihood. Recall from Section 1.1 that the LL estimates are obtained using only the n^i observations in region R^i. When the sample size n^i is fixed and finite, the LL estimate $\hat{\theta}_{LL}^i$ is not consistent for $i = 1, \ldots, I$ and remains inconsistent as I increases.

Let the positive definite Fisher information matrix for θ^i be

$$I_\theta^i = \begin{pmatrix} I_{\psi^i \psi}^i & I_{\psi^i \lambda}^i \\ I_{\lambda^i \psi}^i & I_{\lambda^i \lambda}^i \end{pmatrix}.$$

We assume that the derivatives of $\alpha(\xi; \theta)$ of orders up to 3 exist and meet certain boundedness conditions (Lehmann 1983). By allowing $n^i \to \infty$, the LL estimate $\hat{\theta}_{LL}^i = (\hat{\psi}_{LL}^i, \hat{\lambda}_{LL}^i)$ is asymptotically normal; that is,

$$\sqrt{n^i}(\hat{\theta}_{LL}^i - \theta^i) = \sqrt{n^i} \begin{pmatrix} \hat{\psi}_{LL}^i - \psi^i \\ \hat{\lambda}_{LL}^i - \lambda^i \end{pmatrix} \xrightarrow{L} N(0, I_{\theta^i}^{-1}). \quad (7)$$

For the MSE of the LL estimate $\hat{\theta}_{LL}^i$, we have $n^i \cdot \text{MSE}(\hat{\theta}_{LL}^i) \longrightarrow \text{tr}[(I_{\theta^i}^{-1})]$ as $n^i \to \infty$, where $\text{tr}(A)$ is the trace of a square matrix A, the sum of its diagonal elements. In particular, $n^i \cdot \text{MSE}(\hat{\lambda}_{LL}^i) \longrightarrow \text{tr}[(I_{\theta^i}^{-1})_{\lambda^i \lambda^i}]$, where $(I_{\theta^i}^{-1})_{\lambda^i \lambda^i}$ is the $(m_i - 1) \times (m_i - 1)$ lower right submatrix of $I_{\theta^i}^{-1}$.

2.1.2 Borrowed Strength Profile Likelihood. For the BSPL estimate, denote by

$$I_{\theta^0} = \begin{pmatrix} I_{\psi^0 \psi^0} & I_{\psi^0 \lambda^0} \\ I_{\lambda^0 \psi^0} & I_{\lambda^0 \lambda^0} \end{pmatrix}$$

the Fisher information matrix of the unconditional density $\alpha(\xi; \theta^0)$, where $\theta^0 = (\psi^0, \lambda^0)$. We assume that the assumptions mentioned in Section 2.1.1 for $\alpha(\xi; \theta)$ also holds for $\alpha(\xi, \theta^0)$. In this case, the distribution of $\hat{\psi}_{PL}^0$ is asymptotic normal:

$$\sqrt{n^0}(\hat{\psi}_{PL}^0 - \psi^0) \xrightarrow{L} N\left(0, (I_{\theta^0})_{\psi^0 \psi^0}^{-1}\right) \quad \text{as } I \to \infty, \quad (8)$$

where $(I_{\theta^0})_{\psi^0 \psi^0}^{-1}$ is the $2m \times 2m$ principal submatrix of $(I_{\theta^0})^{-1}$. Equations (7) and (8) indicate that the convergence rate for $\hat{\psi}_{PL}^0$ and $\hat{\psi}_{LL}^i$ is $\sqrt{n^0}$ and $\sqrt{n^i}$, respectively. This leads to the following comparison:

Proposition 1.
1. $\text{MSE}(\hat{\psi}_{LL}^i) = O(1/n^i)$ and $\text{MSE}(\hat{\psi}_{PL}^0) = O(1/n^0)$. As $I \to \infty$ and $n^i \to \infty$, if ψ^0 and λ^i are nonorthogonal, then
2. $n^i \cdot \text{MSE}(\hat{\lambda}_{PL}^i) < n^i \cdot \text{MSE}(\hat{\lambda}_{LL}^i)$.

Proof. See Appendixes A.1 and A.2.

Both the LL and BSPL can be used on overlapping scan windows. Intuitively, the BSPL estimates are superior to the LL estimates because all the observations are used to estimate the structural parameter ψ^0. Proposition 1 shows this superiority of the BSPL estimates explicitly.

2.2 Comparison Between the BSPL Estimate and the JL Estimate

2.2.1 Asymptotic Bias of $\hat{\psi}_{JL}^0$. In this framework, though the structural parameter vector ψ^0 is the same across regions, the incidental parameter vectors λ^i are not assumed to be the same, and their number increases as the number of scan regions $I \to \infty$. Neyman and Scott (1948), through an example, showed that the MLE of the structural parameter may not be consistent. We generalize this result to the case of JL estimation for finite mixture models in spatial scan analysis.

Theorem 1. Consider general m component mixtures of an absolutely continuous exponential family density function $C(\cdot)$ with $\theta = (\psi_1, \dots, \psi_m, \lambda_1, \dots, \lambda_{m-1})'$, an $(m(k+1)-1)$–dimensional parameter vector, where $\psi_t \in \Psi_t \subset \mathbb{R}^k$ is a k-dimensional parameter for the tth component $C(\cdot; \psi_t)$ in the mixture.

If (1) the structural parameter $\psi = (\psi_1, \dots, \psi_m)'$ and the incidental parameter $\lambda = (\lambda_1, \dots, \lambda_m)'$ are nonorthogonal and (2) the existence of, and certain boundedness conditions on, derivatives of $\alpha(\xi; \theta)$ of orders up to 3 are given and $I(\theta_{\text{true}})$, the Fisher information matrix evaluated at the true θ_{true} is assumed to be well defined and positive definite, then $\hat{\psi}_{\text{JL}}^0$ is *asymptotically biased almost surely.*

Proof. See Appendix A.3.

For m component normal mixtures, when n^i is fixed and $I \to \infty$, let $\{b_{\hat{\mu}_t^0}, b_{\hat{\nu}_t^0}\}_{t=1}^m$ be the asymptotic biases of the estimates of $\psi^0 = (\mu_1^0, \nu_1^0, \dots, \mu_m^0, \nu_m^0)'$. These biases can be obtained through the EM algorithm as

$$E(\hat{\mu}_{t\mathrm{JL}}^0) = \mu_t^0 + b_{\hat{\mu}_t^0} = \frac{E(\sum_{j=1}^{n^i} \xi_j^1 \breve{\rho}_{jt}^1)}{n^i E(\hat{\lambda}_{t\mathrm{JL}}^1)}$$

$$E(\hat{\nu}_{t\mathrm{JL}}^0) = \nu_t^0 + b_{\hat{\nu}_t^0} = \frac{E\left(\sum_{j=1}^{n^i} \breve{\rho}_{jt}^1 (\xi_j^1 - \hat{\mu}_t^0)^2\right)}{n^i E(\hat{\lambda}_{t\mathrm{JL}}^1)}, \qquad (9)$$

where $t = 1, \dots, m$. The biases $\{b_{\hat{\mu}_t^0}, b_{\hat{\nu}_t^0}\}_{t=1}^m$ can be obtained by solving the preceding $2m$ equations once we find the values of the expectations on the right side of (9) through numerical integration. For $m = 2$ and for any fixed n^i, $\{\breve{\rho}_{jt}^1\}_{t=1}^m$ is a sequence of deterministic functions of $\{\xi_j^1\}_{j=1}^{n^i}$ conditional on $\hat{\psi}_{\text{JL}}^0$ and is independent of the choice of the starting point $\psi^{0(0)}$. For $m \geq 3$ and $n^i \geq 2$, $\{\breve{\rho}_{jt}^1\}_{t=1}^m$ depends not only on the observations $\{\xi_j^1\}_{j=1}^{n^i}$ but also on $\psi^{0(0)}$ whenever the JL function (4) has more than one local maxima. This dependence on the starting point for the EM algorithm adds complexity to the calculation of the biases for $m \geq 3$.

2.2.2 AMSE($\hat{\lambda}_{\text{JL}}^i$) versus AMSE($\hat{\lambda}_{PL}^i$).

Once we have the asymptotic bias for $\hat{\psi}_{\text{JL}}^0$, the JL estimate $\hat{\lambda}_{t\mathrm{JL}}^i$ for the incidental parameter λ_t^i can be obtained, alternatively, by maximizing the conditional likelihood function

$$L_c^J(\lambda^i | \hat{\psi}_{\text{JL}}^0) = \prod_{j=1}^{n^i} \sum_{t=1}^m \lambda_t^i \phi(\xi_j^i; \hat{\psi}_{\text{JL}}^0).$$

Define the asymptotic MSE AMSE ($\bar{\lambda}^i$) to be the limit of MSE ($\bar{\lambda}^i$) as $I \to \infty$. Provided that the random variables $\hat{\lambda}_{t\mathrm{JL}}^i$ are independent of the starting point for the EM algorithm, the calculation of AMSE ($\hat{\lambda}_{t\mathrm{JL}}^i$) is straightforward.

As shown in Section 2.1.2, the BSPL estimate $\hat{\psi}_{PL}^0$ for the structural parameter ψ_0 is consistent. This result indicates that, with $b_{\hat{\mu}_t^0} = b_{\hat{\nu}_t^0} = 0$ for $t = 1, \dots, m$, the AMSE ($\hat{\lambda}_{PL}^i$) can be calculated in a similar manner as that of AMSE ($\hat{\lambda}_{\text{JL}}^i$) for fixed n^i when $I \to \infty$.

We now consider the simple case of $m = 2$ component normal mixtures with $\nu_1^0 = \nu_2^0 = 1$ and μ_2^0 known. Thus, μ_1^0 and λ_1^i are the only unknown parameters. From (9) it follows that the asymptotic value of $E(\hat{\mu}_{1\mathrm{JL}}^0)$ can be obtained by numerically

evaluating

$$(n^i E(\hat{\lambda}_{1\mathrm{JL}}^i))^{-1} \int_{\mathbb{R}^{n^i}} \left[\sum_{j=1}^{n^i} \xi_j^1 \breve{\rho}_{j1} \right] \prod_{j=1}^{n^i} \left[\sum_{t=1}^m \lambda_t^1 \phi(\xi_j^1; \mu_t^0, \nu_t^0) \right] d\Xi^1,$$

where the integration is multidimensional when $n^i \geq 2$ (see Appendix A.3).

For example, for $n^i = 2$ for all i, the asymptotic bias of $\hat{\mu}_{1\mathrm{JL}}^0$, denoted $b_{\hat{\mu}_{1\mathrm{JL}}^0}(\lambda_1^i)$ to make explicit that this bias is a function of the mixing coefficient, satisfies the following equation:

$$E(\hat{\mu}_{1\mathrm{JL}}^0) = \mu_1^0 + b_{\hat{\mu}_{1\mathrm{JL}}^0}(\lambda_1^i) = \frac{E(\xi_1^1 \breve{\rho}_{j1}^1 + \xi_2^1 \breve{\rho}_{j2}^1)}{2E(\hat{\lambda}_{1\mathrm{JL}}^1)}$$

$$= \frac{\iint_{\mathbb{R}^2} (\xi_1^1 \breve{\rho}_{j1}^1 (\xi_1^1, \xi_2^1, E(\hat{\mu}_{1\mathrm{JL}}^0)) + \xi_2^1 \breve{\rho}_{j2}^1 (\xi_1^1, \xi_2^1, E(\hat{\mu}_{1\mathrm{JL}}^0)) \cdot L(\xi; \theta) d\xi_1^1 d\xi_2^1}{\iint_{\mathbb{R}^2} (\breve{\rho}_{j1}^1 (\xi_1^1, \xi_2^1, E(\hat{\mu}_{1\mathrm{JL}}^0)) + \breve{\rho}_{j2}^1 (\xi_1^1, \xi_2^1, E(\hat{\mu}_{1\mathrm{JL}}^0)) \cdot L(\xi; \theta) d\xi_1^1 d\xi_2^1}, \quad (10)$$

where $L(\xi; \theta) = \prod_{j=1}^2 \left(\sum_{t=1}^2 \lambda_t^1 \phi(\xi_j^1; \mu_t^0, \nu_t^0) \right)$. We calculate $b_{\hat{\mu}_{1\mathrm{JL}}^0}(\lambda_1^i)$, a function of λ_1^i and the difference of component means for the case in which the true $\mu_1^0 = 0$. Figure 1 shows the asymptotic bias of the JL estimate $\hat{\mu}_{1\mathrm{JL}}^0$ from Equation (10) when $\mu_2^0 = 2$ for $0 < \lambda_1^i < 1$, along with Monte Carlo simulation results when the number of regions $I = 100,000$. The figure also depicts the analogous results for the asymptotically unbiased BSPL estimate $\hat{\mu}_{1PL}^0$. The asymptotic bias of $\hat{\mu}_{1\mathrm{JL}}^0$ is strictly decreasing as a function of λ_1^i; thus there is only one choice of λ_1^i for which $b_{\hat{\mu}_{1\mathrm{JL}}^0}(\lambda_1^i) = 0$. That is, $\hat{\mu}_{1\mathrm{JL}}^0$ is asymptotically biased almost everywhere with respect to Lebesgue measure on λ_1^i.

Given the asymptotic bias of $\hat{\mu}_{1\mathrm{JL}}^0$, we also calculate the difference of the AMSE of the BSPL estimate $\hat{\lambda}_{PL}^i$ and the JL estimate $\hat{\lambda}_{\text{JL}}^i$, where $\lambda^i = \lambda_1^i$. Both the theoretical curve and the Monte Carlo simulation results are shown in Figure 2. This figure shows that the improvement in estimation accuracy obtained by using BSPL rather than JL can be significant indeed. It is noteworthy that there is a range of mixing coefficients for which the AMSE of the JL estimate is less
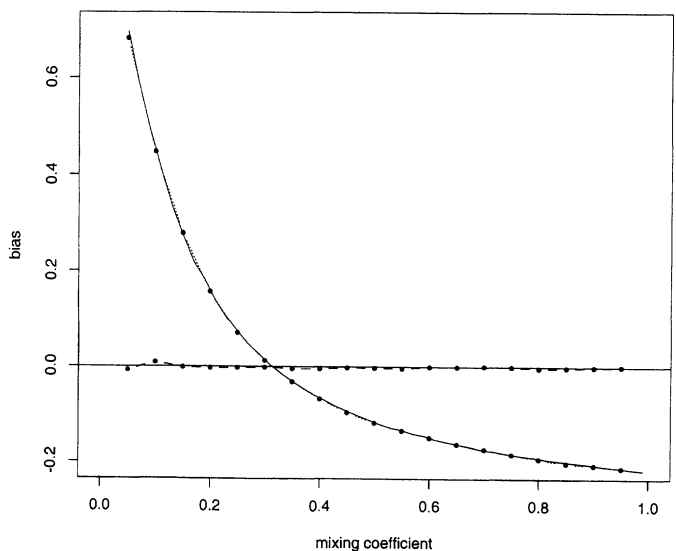

Figure 1. *Theoretical (curve) and Simulation (dots) Results for the Asymptotic Bias of the JL Estimate $\hat{\mu}_{1JL}^0$ and the BSPL Estimate $\hat{\mu}_{1PL}^0$ in Two-Component Normal Mixtures, as a Function of the Mixing Coefficient λ_1^i for $n^i = 2$, $\mu_1^0 = 0$, $\mu_2^0 = 2$, and $\nu_1^0 = \nu_2^0 = 1$:* ——— , *Theoretical JL;* ⋯⋯ , *Simulation JL;* –––, *Simulation PL.*
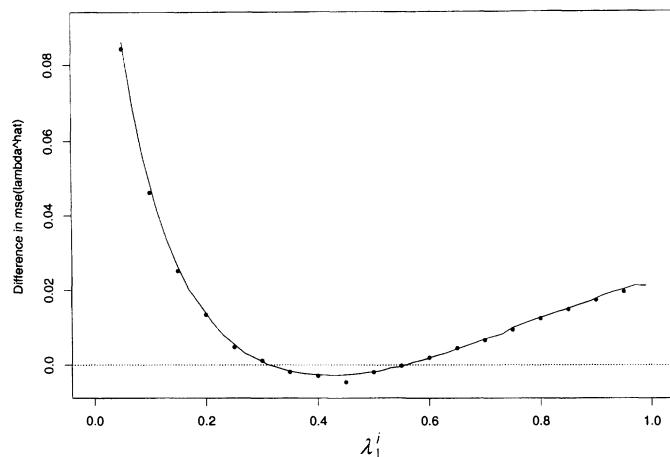
Figure 2. Theoretical (curve) and Simulation (dots) Results for the Difference $D(\hat{\lambda}_{JL}^i, \hat{\lambda}_{PL}^i) = AMSE(\hat{\lambda}_{JL}^i) - AMSE(\hat{\lambda}_{PL}^i)$ of Asymptotic MSE's for the JL Estimate $\hat{\lambda}_{JL}^i$ and the BSPL Estimate $\hat{\lambda}_{PL}^i$, Where $\lambda^i = \lambda_1^i$ in Two-Component Normal Mixtures. It is a function of the mixing coefficient λ_1^i for $n^i = 2$, $\mu_1^0 = 0$, $\mu_2^0 = 2$, and $\nu_1^0 = \nu_2^0 = 1$.

Figure 3. Simulation Results for the Difference $\widehat{D}(\hat{\lambda}_{JL}^1, \hat{\lambda}_{PL}^1) = MSE(\hat{\lambda}_{JL}^1) - MSE(\hat{\lambda}_{PL}^1)$ of the MSE for the JL estimate $\hat{\lambda}_{JL}^1$ and the BSPL estimate $\hat{\lambda}_{PL}^1$, where $\lambda^1 = \lambda_1^1$, as a function of μ_2^0 for $I = 100$, $n^1 = 100$, $\mu_1^0 = 0$, $\lambda_1^1 = .1$, and $\nu_1^0 = \nu_2^0 = 1$.

than that of the BSPL estimate. [The theoretical curve for $D(\hat{\lambda}_{JL}^i, \hat{\lambda}_{PL}^i) = AMSE(\hat{\lambda}_{JL}^i) - AMSE(\hat{\lambda}_{PL}^i)$ does fall slightly below 0 for a small range $(\lambda_u, \lambda_u + \delta)$, where λ_u is such that $b_{\hat{\mu}_1^0}(\lambda_u) = 0$.] Examination of Figure 2 indicates that the minimax rule favors the BSPL estimate; when the AMSE of $\hat{\lambda}_{JL}^i$ is smaller than that of $\hat{\lambda}_{PL}^i$, the difference is negligible compared to the difference of the two MSE's when the BSPL estimate is better. That is, the BSPL estimate $\hat{\theta}_{PL}^i$ is *minimax* over $C = \{\hat{\theta}_{PL}^i, \hat{\theta}_{JL}^i\}$, meaning $\max_{0 < \lambda_1^i < 1} D(\hat{\lambda}_{PL}^i, \hat{\lambda}_{JL}^i) < \max_{0 < \lambda_1^i < 1} D(\hat{\lambda}_{JL}^i, \hat{\lambda}_{PL}^i)$ for $i = 1, \ldots, I$. [In general, $\hat{\theta}$ is minimax over $C \Longleftrightarrow$

$$\hat{\theta} = \operatorname*{argmin!}_{\hat{\theta}'' \in C} \max_{\hat{\theta}' \in C \setminus \hat{\theta}''} \max_{0 < \lambda < 1} D(\hat{\lambda}'', \hat{\lambda}'),$$

where ! stands for the uniqueness of the minimum.]

In fact, having demonstrated our minimax superiority of BSPL for $\theta^* \in \{(\mu_1^0, \nu_1^0, \mu_2^0, \nu_2^0, \lambda_1^i) : \mu_2^0 - \mu_1^0 = 2, \nu_1^0 = \nu_2^0 = 1,$ and $0 < \lambda_1^i < 1\}$ and appealing to continuity of AMSE as a function of λ, we have the following theorem.

Theorem 2. BSPL is superior to JL in our minimax sense in an open neighborhood $\Theta_{\text{minimax}} \subset \Theta$ around θ^*.

General analytic results analogous to those just presented require significant computational effort. In particular, the multidimensional integration required for the asymptotic value of $E(\hat{\mu}_{rJL}^0)$ becomes unwieldy for large values of n^i. (The almost sure asymptotic biasedness of the JL estimate from Theorem 1 holds; however, obtaining the exact value for this bias is a computational issue.)

To demonstrate that the minimax superiority of BSPL generalizes to large values of n^i, a Monte Carlo simulation is presented. We choose $I = 100$ regions and generate $n^i = 100$ observations in each region. The observations in the first 10 regions have the underlying marginal density $\alpha(\xi; \theta^1) = .1\phi(\xi; 0, 1) + .9\phi(\xi; \mu_2^0, 1)$, whereas the observations in the remaining 90 regions have the underlying marginal density $\alpha(\xi; \theta^2) = .54\phi(\xi; 0, 1) + .46\phi(\xi; \mu_2^0, 1)$. The unconditional density is therefore $\alpha(\xi; \theta^0) = .5\phi(\xi; 0, 1) + .5\phi(\xi; \mu_2^0, 1)$.
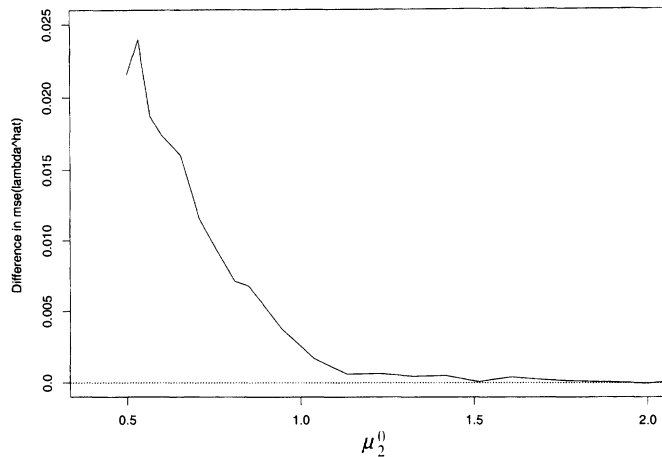
We consider estimating $\theta^1 = (0, 1, \mu_2^0, 1, .1)$. For various values of μ_2^0 we perform 100 Monte Carlo replicates for each, obtaining estimates $\hat{\theta}_{PL}^1$ and $\hat{\theta}_{JL}^1$ for the underlying density $\alpha(\xi; \theta^1)$ of the first region. Figure 3 shows the difference $\widehat{D}(\hat{\lambda}_{JL}^1, \hat{\lambda}_{PL}^1) = \widehat{MSE}(\hat{\lambda}_{JL}^1) - \widehat{MSE}(\hat{\lambda}_{PL}^1)$ between the estimated MSE's of $\hat{\theta}_{PL}^1$ and $\hat{\theta}_{JL}^1$ as a function of μ_2^0. The simulation results indicate the minimax superiority of the BSPL estimate $\hat{\theta}_{PL}^1$ for $\theta^* \in \{(\mu_1^0, \nu_1^0, \mu_2^0, \nu_2^0, \lambda_1^1) : 0 < \mu_2^0 - \mu_1^0 < \infty, \nu_1^0 = \nu_2^0 = 1, \lambda_1^1 = .1\}$. A sign test for $H_0 : \text{median}(\widehat{MSE}(\hat{\lambda}_{JL}^1) - \widehat{MSE}(\hat{\lambda}_{PL}^1)) \leq 0$ results in a p value of .003 at $\mu_2^0 - \mu_1^0 = .8$, providing an indication of the statistical significance of the BSPL superiority represented in Figure 3.

2.3 The Choice of Distance Measure

Given the similarity assumption on the underlying marginal density α^i proposed in Section 1—that is, different scan regions share the same structural parameter ψ^0 and the difference lies entirely in the incidental parameters (mixing coefficients) λ^i—it seems natural to use a measure based only on the incidental parameters. That is, consider $\text{se}(\lambda^i, \lambda^j) = (\lambda^i - \lambda^j)'(\lambda^i - \lambda^j)$. As discussed in the introduction, a better estimate of $\hat{\lambda}^i$, in terms of a smaller MSE, will increase the power of the test. The question is whether the test based on SE detects the nonhomogeneity of interest in the application. For example, suppose we have, as the underlying marginal density for the background region, the three-component normal mixture given by $\alpha^0(x) = .76\phi(x; 0, 1) + .05\phi(x; 1, .5) + .19\phi(x; 1.25, .0625)$ and the marginal density for the ROI region has the form $\alpha^1(x) = .4\phi(x; 0, 1) + .1\phi(x; 1, .5) + .5\phi(x; 1.25, .0625)$. Suppose we have two estimates for α^1, denoted $\hat{\alpha}_a^1$ and $\hat{\alpha}_b^1$, given by $\hat{\alpha}_a^1(x) = .42\phi(x; 0, 1) + .11\phi(x; 1, .5) + .47\phi(x; 1.25, .0625)$ and $\hat{\alpha}_b^1(x) = .50\phi(x; 0, 1) + .15\phi(x; 1, .5) + .35\phi(x; 1.25, .0625)$. The density estimate $\hat{\alpha}_a^1$ is superior to $\hat{\alpha}_b^1$ in terms of SE of the mixing coefficients and ISE. Figure 4 plots the four densities $\alpha^0, \alpha^1, \hat{\alpha}_a^1$, and $\hat{\alpha}_b^1$ together, illustrating that SE is an appropriate measure of distance for this case in that the test based on SE is more reliable when using $\hat{\alpha}_a^1$, as desired.
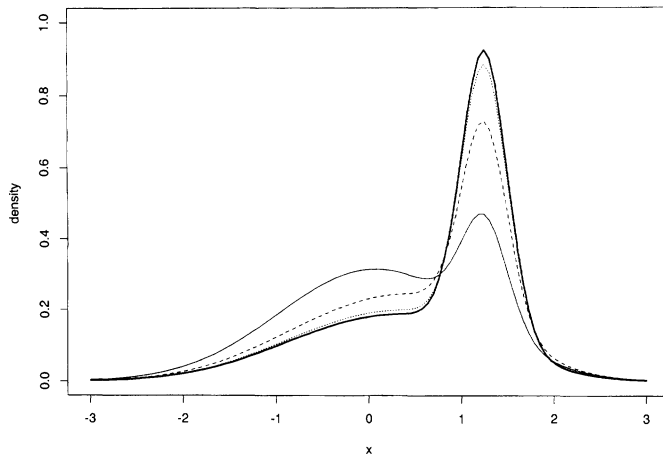
Figure 4. This Plot Gives an Example for Which the SE of the Mixing Coefficients Is an Appropriate Distance Measure for Finite Mixture Models. The two solid curves (α^0 and α^1) are the underlying marginal densities for background and ROI regions (bold), respectively. The dotted curve ($\hat{\alpha}_a^1$) and the dashed curve ($\hat{\alpha}_b^1$) are two density estimates for α^1. The density estimate $\hat{\alpha}_a^1$ has a smaller SE, and it is clear that $\hat{\alpha}_a^1$ is superior to $\hat{\alpha}_b^1$ in terms of ISE as well. This illustrates that SE is an appropriate measure of distance in this case.

Unfortunately, SE is not a universally appropriate measure. For example, for finite normal mixtures with more than two components, when the variances of the normal components are of the same scale and two of the means are close, the SE of the mixing coefficient λ^i may not capture the important difference between densities. It can be seen from this example that finding better parameter estimates and choosing an appropriate measure of distance are two separate issues. Although better parameter estimates yield higher power in the test of nonhomogeneity, an inappropriate measure can lead to an incorrect conclusion from the test. To demonstrate, suppose that the baseline density function is given by

$$\alpha^0(x) = \frac{1}{3}\phi(x; 0, 1) + \frac{1}{3}\phi(x; .01, 1) + \frac{1}{3}\phi(x; 10, 1).$$

Then clearly (see Fig. 5)

$$\alpha^1(x) = \frac{2}{3}\phi(x; 0, 1) + \frac{1}{3}\phi(x; 10, 1)$$

is much closer to α^0 than

$$\alpha^2(x) = \frac{1}{2}\phi(x; 0, 1) + \frac{1}{2}\phi(x; .01, 1)$$

in terms of the ISE. SE indicates, however, that α^2 is closer to α^0.

Under these circumstances, the ISE

$$\text{ISE}(\hat{\theta}^i) = \int (\alpha^i(\xi; \hat{\theta}^i) - \alpha(\xi; \theta))^2 d\xi$$

is a better choice than $\text{SE}(\hat{\lambda}^i)$. A quadratic approximation relating $\text{MSE}(\hat{\psi}^0)$ and $\text{MSE}(\hat{\lambda}^i)$ with $E[\text{ISE}(\hat{\theta}^i)]$ can be established via Taylor expansion. Intuitively, for large sample sizes (when all estimates are close to the true parameters) a smaller $\text{MSE}(\hat{\psi}^0)$ and $\text{MSE}(\hat{\lambda}^i)$ yields a smaller $E[\text{ISE}(\hat{\theta}^i)]$. Due to the limitations of $\text{SE}(\hat{\lambda}^i)$ for small samples (when n^i is small), ROI identification results in Section 3 are presented using both SE and ISE. Although other functional measures, such as integrated absolute error distance L, might
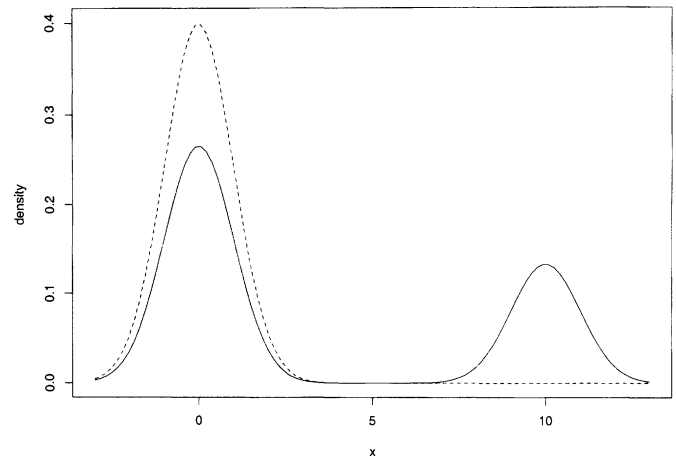

Figure 5. This Figure Gives an Example of the Limitations of Using the SE of the Mixing Coefficients. The density α^0 (solid curve) and α^1 are essentially identical (indistinguishable in the plot). The SE between α^0 and α^1, however, is larger than the SE between the density α^2 (dashed curve) and α^0.

provide more suitable distance metrics for this application (see Devroye and Gyorfi 1985), we chose the ISE distance because of its computational advantages, (closed form for mixtures of normals).

3. APPLICATION

We now describe the use of borrowed strength finite mixture estimates in spatial scan analysis for ROI identification. For illustration we consider the identification of ROI in unmanned aerial vehicle (UAV) imagery. The infrared imagery is extracted from a videotape collected from a UAV test flight over the Naval Strike Warfare Center, Fallon, Nevada, in the summer of 1995. Figure 6 presents one image from the Fallon videotape. Previous results on imagery extracted from this videotape were given by Solka et al. (1997a,b, 1998). Although the buildings in the image are easy to observe, detecting the track made by vehicle traffic in the upper left region of the image is a more challenging problem. We consider as our domain of interest R^0 as annotated in Figure 6. Many realistic ROI identification scenarios have multiple target and/or background types. Preliminary results for the Fallon UAV imagery using the BSPL approach were given by Priebe, Solka, and Tao (1997), who compared the regional estimates $\hat{\alpha}_{\text{PL}}^i$ against estimates of each background model α^{B_l} in the class $A = \{\alpha^{B_1}, \ldots, \alpha^{B_L}\}$. Such a procedure allows for detecting ROI in a more cluttered scene but requires estimation of each α^{B_l}. For simplicity, we have chosen a domain of interest involving only one "background" and one "target."

For many detection problems such as this, the pixel intensity observations are not particularly useful. Often consideration of a local "texture," or roughness, feature can improve detection capability. We investigate the marginal probability density function of derived texture features $\zeta : \zeta(\mathbf{x}) = h(\{\xi(\mathbf{y}) : \mathbf{y} \in B(\mathbf{x}, \epsilon)\})$, where $h(\cdot)$ is a function defined on a neighborhood $B(\mathbf{x}, \epsilon)$, $\epsilon > 0$ for each spatial location $\mathbf{x}$. An example of a texture feature that characterizes local roughness is the coefficient of variation; $\zeta(\mathbf{x}) = S(\mathbf{x})/\bar{\xi}(\mathbf{x})$, where the sample statistics $\bar{\xi}(\mathbf{x}) = 1/|B| \sum_{\mathbf{y} \in B(\mathbf{x}, \epsilon)} \xi(\mathbf{y})$ and $S(\mathbf{x}) = 1/|B| \sum_{\mathbf{y} \in B(\mathbf{x}, \epsilon)} (\xi(\mathbf{y}) -$
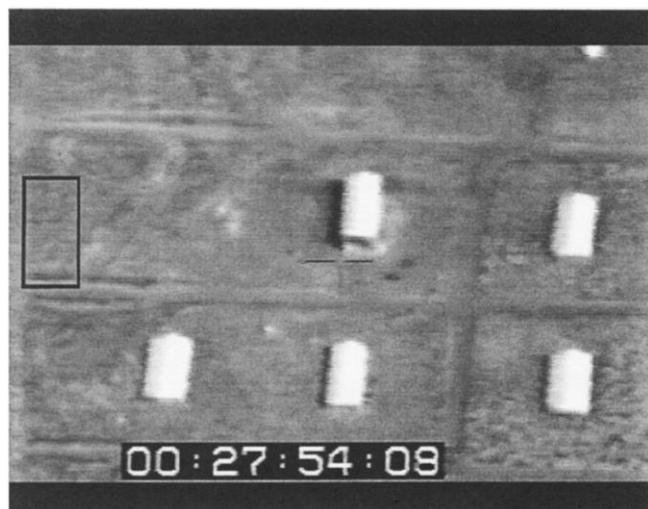
Figure 6. One Image From the Fallon Videotape. The rectangular region in the upper left is the target area R^0. The track made by vehicle traffic running horizontally across the bottom of R^0 is the ROI.

$\bar{\xi}(\mathbf{x}))^2$ are obtained based on the gray-level pixel intensity observations in $B(\mathbf{x}, \epsilon)$. $B(\mathbf{x}, \epsilon) = \{\mathbf{y} : ||\mathbf{y} - \mathbf{x}||_2 \leq \epsilon\}$, and we choose $\epsilon = 3$ so that $B(\mathbf{x}, \epsilon)$ contains the observations in the circular area centered at $\mathbf{x}$ with radius 3. The realization of coefficient of variation field ζ is given in Figure 7.

[The coefficient of variation is chosen for simplicity. Feature extraction for ATR imagery is an active area of investigation in science and engineering. In fact, because multivariate features

Figure 7. The Realization of the Coefficient of Variation Field ζ.

are commonly extracted in image analysis, we note that the extension of the borrowed strength methodology to multivariate observations $\zeta(\mathbf{x})$ is straightforward. Although the scan regions R^i are small, only the mixing coefficients are estimated for each. Thus, the dimensionality of the mixture parameter space comes into play only when estimating α^0, for which a large sample is available. For an investigation of local feature extraction and a comparison of coefficient of variation with a more elaborate multivariate feature, see Marchette, Lorey, and Priebe (1997).]

Figure 8 depicts the directional correlogram (Cressie 1993) of ζ for a subregion of background. The assumptions of stationarity and isotropy in the coefficient-of-variation field for a background desert patch are both reasonable and consistent with this directional correlogram. The dependence structure is such that a scan region in ζ containing 64 observations has an effective sample size of approximately 10 independent observations; the effective sample sizes are indeed small.

Given the coefficient-of-variation observations in the overall region R^0, an unconditional finite mixture density estimate $\hat{\alpha}^0 = \alpha(\xi; \hat{\psi}_{PL}^0, \hat{\lambda}^0)$ is obtained using all n^0 available observations. This estimate also requires estimating the number of terms m in this overall mixture. The theoretical results presented are conditional on the estimate $\hat{m}$ being the true order of the mixture. There is a wealth of literature studying the automatic determination of the number of components to be used in a finite mixture model. We follow the procedure of Priebe and Marchette (2000) using an alternating kernel and mixture (AKM) algorithm in which the number of components is determined by increasing the complexity of the mixture model based on mismatch between the current mixture estimate and a (filtered) kernel estimate. The estimate $\hat{\alpha}^0$ fixes $\hat{m}$ and the structural parameters ψ^0 at $\hat{\psi}_{PL}^0$ for the subsequent spatial scan analysis. Figure 9 presents the three-component finite mixture model $\hat{\alpha}^0$ produced by AKM, overlaid with the histogram of coefficient-of-variation observations. By increasing the number of components in the mixture model, one can improve the goodness of fit of the overall density estimate $\hat{\alpha}^0$ but at a cost of additional parameters that adversely affects

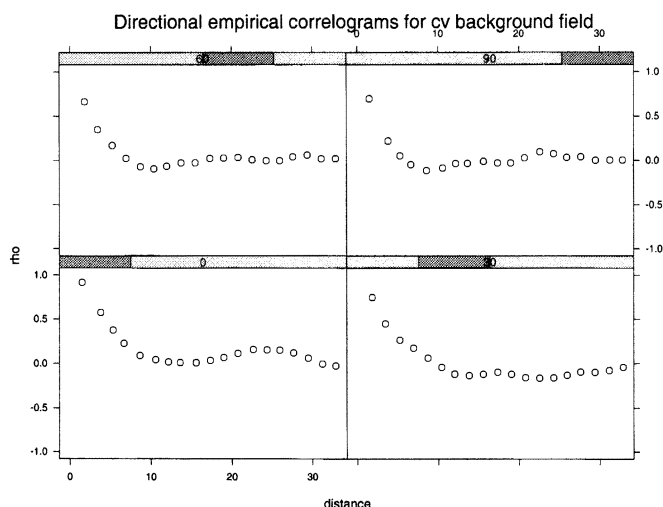Directional empirical correlograms for cv background field

Figure 8. The Directional Correlogram of the Coefficient-of-Variation Field ζ for a Subregion of Background. Note that ζ appears stationary and isotropic with effective sample-size factor of .15.
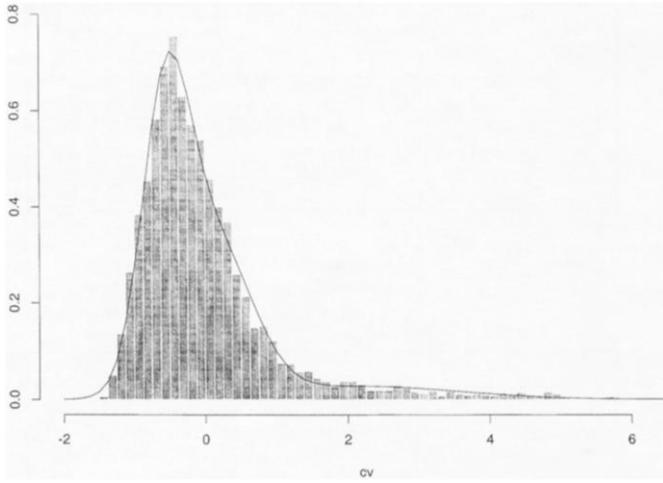
Figure 9. The Three-Component Finite Mixture Model $\widehat{\alpha}^0$ Produced by AKM, Overlaid With Histogram of Coefficient-of-Variation Texture Features $\zeta(x)$ for $x \in R^0$.

the small-sample local estimates. This trade-off between accuracy of $\widehat{\alpha}^0$ and complexity of local estimates is fundamental. A robustness analysis, addressing model misspecification, is an obvious candidate for further investigation.

Due to the restriction of the JL method to overlapping scan regions, we present the comparison of BSPL with JL in this suboptimal framework. (Note, however, that an additional aspect favoring BSPL is its applicability to overlapping scan regions, which are superior in practice because they eliminate boundary issues in detection.) We divide R^0 into 72 disjoint squared regions, each containing $8 \times 8 = 64$ observations. The JL estimates $\widehat{\psi}_{JL}^0$ and $(\widehat{\lambda}_{JL}^1, \ldots, \widehat{\lambda}_{JL}^{72})$ are obtained simultaneously through the JL (4). From this we obtain $\widehat{\alpha}_{JL}^i = \alpha(\xi; \widehat{\psi}_{JL}^0, \widehat{\lambda}_{JL}^i)$.

For each of the 72 scan regions R^i, a profile estimate of the mixing coefficients $\widehat{\lambda}_{PL}^i$ is obtained based on α^0. This part of the algorithm—the time-critical part for many applications in which $\widehat{\alpha}^0$ can be fixed prior to local estimation—is not computationally demanding. The computational resources necessary for these spatial scan density estimates can be further reduced by employing a modification of the EM algorithm specific to mixtures in which only the mixing coefficients need to be estimated (Pilla 1997).

As noted previously, for this application the majority of R^0 consists of "background"; the alternative hypothesis of interest is that most of R^0 has marginal α^B, and there is only a small ROI with marginal α^T. Thus, the overall marginal density α^0 is close to that of the background; $\alpha^0 \approx \alpha^B$. Given an estimate $\widehat{\alpha}^0$ of α^B of the marginal density of background, we consider statistics $\widehat{T}^i = d(\widehat{\alpha}^i, \widehat{\alpha}^0) = \text{SE}(\widehat{\lambda}^i, \widehat{\lambda}^0)$. Then $\widehat{T} = \max_{i=1,\ldots,I} \widehat{T}^i$ is the test statistic for our test of nonhomogeneity, and we reject for large values of $\widehat{T}$. We consider deviations from $\widehat{\alpha}^0$ as our candidate ROI; our interest is in those scan regions R^i such that $\widehat{T}_{PL}^i \geq \tau_{PL}$ or $\widehat{T}_{JL}^i \geq \tau_{JL}$ for BSPL or JL, respectively. For quatitative statistical inference, it remains only to determine the threshhold τ. Determination of τ for a given significance level is a nontrivial multiple comparisons problem, however.

When the sample size n^i is large, the field formed by the test statistics $\widehat{T}_{PL}^i$ (appropriately normalized) based on the BSPL

estimate is approximately a χ^2 field. In this case, one can determine an appropriate critical value via either the Euler characteristic theory of Worsley (1994) or importance sampling (Naiman and Priebe in press).

The small effective sample sizes available for the regional mixing coefficient estimates make appeals to asymptotic theory unattractive. Furthermore, the process of utilizing a post-processing step in which candidate ROI are analyzed for spatial clustering is appropriate against the class of alternatives for which R^0 is partitioned into background and target with the target consisting of a small, contiguous anomaly. Because ROI identification is but the first stage in a full ATR system, post-processing steps will be put in place, and our consideration of post-processing for candidate ROI is not unduly burdensome. Our approach is to analyze the scan regions corresponding to, say, the largest $k\%$ of the $\widehat{T}^i$ for spatial clustering. If many of these anomalous scan regions are clustered in R^0, a potential ROI for further investigation has been identified. Under H_0 these anomalous scan regions are expected to be approximately uniformly distributed in R^0. (Dependency considerations, both in components of image observations and overlapping scan regions, complicate this issue.) The theoretical results from Section 2.2 suggest that, for large I, testing using $\widehat{T}_{PL}^i$ can perform better than testing using $\widehat{T}_{JL}^i$.

Thus, for the BSPL UAV ROI, we have developed the following algorithm in automatic target recognition:

Step 1: Determine the scan regions R^i.
Step 2: Obtain $\widehat{m}$ and $\widehat{\alpha}^0 = \alpha(\widehat{\theta}_{PL}^0) = \alpha(\widehat{\psi}_{PL}^0, \widehat{\lambda}^0)$ for R^0.
Step 3: Obtain the local estimates $\widehat{\lambda}_{PL}^i$ for $i = 1, \ldots, I$.
Step 4: Calculate $\widehat{T}_{PL}^i$ for $i = 1, \ldots, I$.
Step 5: Determine the location of candidate ROI as the largest $k\%$ of the $\widehat{T}_{PL}^i$.
Step 6: Analyze candidate ROI for spatial clustering.

[Note that for Step 1 the choice of scan regions as balls or rectangles is artificial. In practice, image segmentation can be employed to partition R^0 into a disjoint union of subregions. These subregions can then be used as the scan regions, allowing for the incorporation of edge information into the process, with the price being the added complexity of unequal and random sample sizes n^i. This approach was investigated for conventional spatial scan analysis of point processes by Priebe, Olson, and Healy (1997) and for the detection of tumorous masses in digital mammography via borrowed strength by Priebe, Marchette, and Rogers (1997).]

For our example from the Fallon video, Figure 10 shows the plot of $\widehat{T}_{PL}^i$ and $\widehat{T}_{JL}^i$ as well as the analogous ISE plots for the selection of 72 square scan regions. The BSPL procedure yields candidate ROI that cluster spatially on the track. Post-processing these candidate ROI yields detection of the track as a region of interest. The candidate ROI obtained using the analogous JL algorithm are less accurate. Although the superiority of BSPL is more dramatic using ISE, the BSPL detections are more reliable then those obtained from JL when using SE as well.

4. DISCUSSION

We have presented a BSPL approach to spatial scan density estimation that is applicable to detecting candidate regions
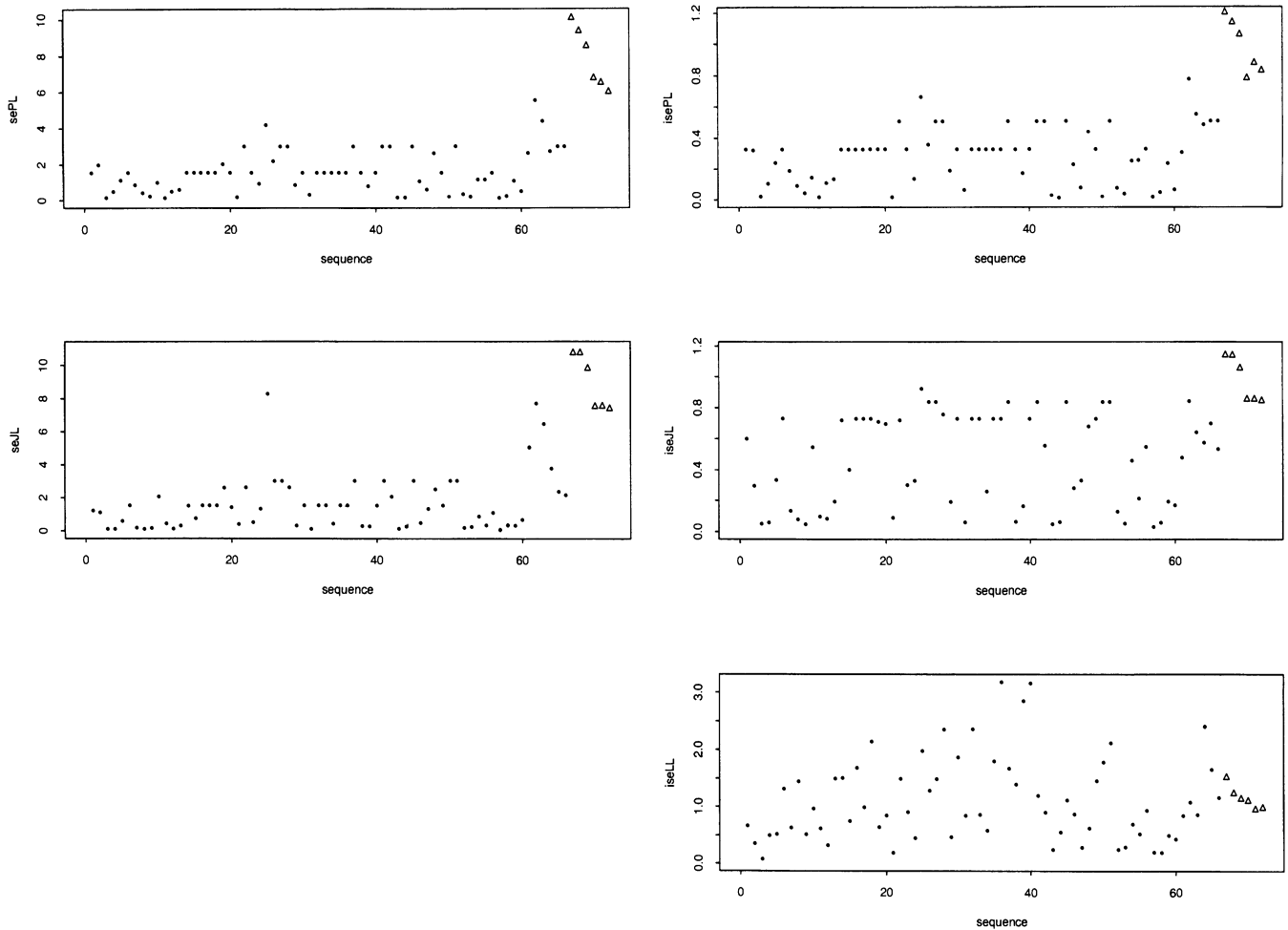
Figure 10. The Plot of the $\widehat{T}_{PL}^j$ and $\widehat{T}_{JL}^j$, as Well as the Analogous ISE Results, for 72 Scan Regions. The triangles represent the track in the region.

of interest for further investigation. Comparisons with traditional likelihood methodologies have been carried out for the realistic scenario in which the number of regions $I \to \infty$. Comparing the BSPL with the LL approach, we find that the BSPL estimate is superior in terms of the MSE. For JL, a generalization of the result of Neyman and Scott (1948) shows that the JL estimate is asymptotically biased almost surely. From this we obtain the superiority of the BSPL estimate over the JL estimate in terms of the MSE in a minimax sense.

The BSPL methodology is applied to the problem of detecting candidate regions of interest in UAV imagery. The theoretical superiority of the BSPL estimate is supported by example results that suggest that, for this problem, the candidate detections obtained via the BSPL approach are more valuable for detecting spatial anomalies than those obtained via JL. We consider both SE of the parameter vector and ISE in the application due to the potential limitations of the former. The fact that these two measures of distance yield similar detection results suggests that the theoretical developments presented in Section 2 for SE are indeed relevant for our application. We conclude that the superiority of the BSPL estimate combined with an appropriate choice for the measure of distance yields the superior performance of the BSPL methodology.

In summary, spatial scan density estimation for finite mixture models (with fixed regional sample sizes and the

number of such regions increasing to infinity) is precisely the application framework for which a generalization of the Neyman and Scott (1948) example, based on the idea of nonorthogonality, indicates that an alternative estimation technique to JL is required. Profile likelihood is a promising approach to this estimation problem and is applicable in ATR. One aspect of a successful ATR system will be ROI identification based on low-level image characteristics, and BSPL offers a contribution.

ACKNOWLEDGMENTS

APPENDIX: PROOFS

A.1   Proof of Part 1 of Proposition 1

Lemma. Let

$$A = \begin{pmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{pmatrix}$$

be positive definite, where A_{00} and A_{11} are square submatrices of A and $A_{01} \neq 0$. Denote

$$A^{-1} = \begin{pmatrix} B_{00} & B_{01} \\ B_{10} & B_{11} \end{pmatrix},$$

where the corresponding B_{ij}'s have the same order as A_{ij}'s. Then we have

$$\mathrm{tr}(B_{11}) > \mathrm{tr}(A_{11}^{-1}).$$

Proof of the Lemma. It is easy to see that $A_{11}^{-1} = B_{11} - B_{10}B_{00}B_{01}$. Because $B_{10} \neq 0$ and B_{00} is positive definite, we obtain $\mathrm{tr}(B_{10}B_{00}B_{01}) = \mathrm{tr}(B_{11} - A_{11}^{-1}) > 0$; thus $\mathrm{tr}(B_{11}) > \mathrm{tr}(A_{11}^{-1})$.

A.2 Proof of Part 2 of Proposition 1

As $I \to \infty$ and $n^i \to \infty$, from the asymptotic normality of $\hat{\boldsymbol{\lambda}}_{\mathrm{PL}}^i$ and $\hat{\boldsymbol{\lambda}}_{\mathrm{LL}}^i$, we have

$$n^i \cdot \mathrm{MSE}(\hat{\boldsymbol{\lambda}}_{\mathrm{LL}}^i) \longrightarrow \mathrm{tr}\left((I_{\theta^i}^{-1})_{\lambda^i \lambda^i} \right)$$

and

$$n^i \cdot \mathrm{MSE}(\hat{\boldsymbol{\lambda}}_{\mathrm{PL}}^i) \longrightarrow \mathrm{tr}\left([(I_{\theta^i})_{\lambda^i \lambda^i}]^{-1} \right),$$

where $(I_{\theta^i})_{\lambda^i \lambda^i}$ is the $(m-1) \times (m-1)$ lower right submatrix of I_{θ^i}.

In the previous lemma, let $A = I_{\theta^i}$, $A_{11} = (I_{\theta^i})_{\lambda^i \lambda^i}$, and $B_{11} = (I_{\theta^i}^{-1})_{\lambda^i \lambda^i}$; then we have

$$\mathrm{tr}\left([(I_{\theta^i})_{\lambda^i \lambda^i}]^{-1} \right) < \mathrm{tr}\left((I_{\theta^i}^{-1})_{\lambda^i \lambda^i} \right).$$

Thus, asymptotically, $n^i \cdot \mathrm{MSE}(\hat{\boldsymbol{\lambda}}_{\mathrm{PL}}^i) < n^i \cdot \mathrm{MSE}(\hat{\boldsymbol{\lambda}}_{\mathrm{LL}}^i)$.

A.3 Proof of Theorem 1

First we define the feasible region of the parameter as $\Theta_F = \{\boldsymbol{\theta} : 0 \leq \lambda_t \leq 1 \text{ for } t = 1, \dots, m-1 \text{ and } \psi_t \in \Psi_t \subset \mathbb{R}^k \text{ for } t = 1, \dots, m\}$ and denote Θ_U to be the set of $\boldsymbol{\theta}$ of which $\hat{\theta}_{\mathrm{JL}}^i$ is asymptotically unbiased. That is, $\Theta_U = \{\boldsymbol{\theta} \in \Theta_F : \hat{\theta}_{\mathrm{JL}}^i \text{ is asymptotically unbiased}\}$.

Moreover, let ν be any "nondegenerate" σ-finite measure on $\mathbb{R}^{m(k+1)-1}$ such that $\nu(\mathcal{M}) = 0$ for all $(m(k+1)-2)$-dimensional manifolds $\mathcal{M}$. One example for this nondegenerate measure is the Lebesque measure on $\mathbb{R}^{m(k+1)-1}$.

To prove Theorem 1, we only need to show that $\nu(\Theta_U) = 0$. We sketch a proof for the case in which the mixing densities $C(\cdot)$ are normal densities. The general case proceeds analogously. If the starting point is "close" to the unique (almost surely) global maxima of the JL function (4), then the EM algorithm converges (almost surely) to the MLE of (4); that is, $\{\check{\lambda}_t^{i(k)}, \check{\mu}_t^{(k)}, \check{\nu}_t^{(k)}\}$ converges to $\{\hat{\lambda}_{iJL}^i, \hat{\mu}_{iJL}, \hat{\nu}_{iJL}\}$ as the number of iterations goes to infinity. Let $\Omega_\mu = \{\boldsymbol{\theta} \in \Theta_F : \hat{\mu}_{1JL}^0 \text{ is unbiased}\}$. Then $\Theta_U \subset \Omega_\mu$. Any $\boldsymbol{\theta}^i \in \Omega_\mu$ satisfies

$$\mu_1^0 = E(\hat{\mu}_{1JL}^0) = E\left[\frac{\frac{1}{n^i}\sum_{j=1}^{n^i} \xi_j^1 \check{\rho}_{j1}^1}{\lim_{I \to \infty} \frac{1}{I}\sum_{i=1}^{I} \hat{\lambda}_{1JL}^i} \right] = \frac{\frac{1}{n^i}E\left[\sum_{j=1}^{n^i} \xi_j^1 \check{\rho}_{j1}^1\right]}{E(\hat{\lambda}_{1JL}^i)}$$

$$= (n^i E(\hat{\lambda}_{1JL}^i))^{-1} \int_{\mathbb{R}^{n^i}} \left[\sum_{j=1}^{n^i} \xi_j^1 \check{\rho}_{j1}^1 \right]$$

$$\times \prod_{j=1}^{n^i} \left[\sum_{t=1}^{m} \lambda_t^1 \phi(\xi_j^1; \mu_t^0, \nu_t^0) \right] d\Xi^1, \qquad (A.1)$$

where $\Xi^1 = (\xi_1^1, \dots, \xi_{n^i}^1)'$.

From Equation (A.1) we see that for any $\boldsymbol{\theta}^i = (\mu_1^0, \nu_1^0, \dots, \mu_m^0, \nu_m^0, \lambda_1^i, \dots \lambda_{m-1}^i)' \in \Omega_\mu$, μ_1^0 can be represented

by an n^ith degree polynomial in each λ_t^i, which implies that the dimensionality of the solution space Ω_μ for (A.1) in terms of μ_1^0 is $\dim(\Omega_\mu) \leq \dim(\Theta_F) - 1 = m(k+1) - 2$.

Thus, $\nu(\Omega_\mu) = 0$ [a straightforward generalization of theorem 12.1 (Billingsley 1995)] and hence $\nu(\Theta_U) = 0$.

REFERENCES

Billingsley, P. (1995), *Probability and Measure*, New York: Wiley.
Cox, D. R., and Reid, N. (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*, New York: Wiley.
Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.
Devroye, L., and Gyorfi, L. (1985), *Nonparametric Density Estimation*, New York: Wiley.
Guyon, X. (1995), *Random Fields on a Network*, New York: Springer-Verlag.
IEEE Transactions on Image Processing (1997), Vol. 6, No. 1, Special Issue on Automatic Target Recognition.
Lehmann, E. L. (1983), *Theory of Point Estimation*, New York: Wiley.
Marchette, D. J., Lorey, R. A., and Priebe, C. E. (1997), "An Analysis of Local Feature Extraction in Digital Mammography," *Pattern Recognition*, 30, 1547–1554.
Marron, J. S., and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics*, 20, 712–736.
McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, New York: Chapman and Hall.
McLachlan, G. J., and Krishnan, T. (1997), *The EM Algorithm and Extensions*, New York: Wiley.
Naiman, D. Q., and Priebe, C. E. (in press), "Computing Scan Statistic p Values Using Importance Sampling, With Applications to Genetics and Medical Image Analysis," *Journal of Computational and Graphical Statistics*, 9.
Neyman, J., and Scott, E. L. (1948), "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16, 1–32.
Pilla, R. S. (1997), "Improving the Rate of Convergence of EM in High-dimensional Finite Mixtures," unpublished Ph.D. dissertation, The Pennsylvania State University, Dept. of Statistics.
Priebe, C. E. (1996), "Nonhomogeneity Analysis Using Borrowed Strength," *Journal of the American Statistical Association*, 91, 1497–1503.
Priebe, C. E., and Marchette, D. J. (2000), "Alternating Kernel and Mixture Density Estimates," *Computational Statistics and Data Analysis*, 35, 43–65.
Priebe, C. E., Marchette, D. J., and Rogers, G. W. (1997), "Segmentation of Random Fields via Borrowed Strength Density Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 494–499.
Priebe, C. E., Olson, T., and Healy, D. M. (1997), "A Spatial Scan Statistic for Stochastic Scan Partitions," *Journal of the American Statistical Association*, 92, 1476–1484.
Priebe, C. E., Solka, J. L., and Tao, P. (1997), "Spatial Scan Analysis of Unmanned Aerial Vehicle Imagery," in *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 92–97.
Redner, R. A., and Walker, H. F. (1984), "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, 26, 195–238.
Solka, J. L., Lorey, R. A., Marchette, D. J., Rogers, G. W., Irwin, V. L., and Wallet, B. C. (1997a), "Evaluation of Multiple Feature Sets for the Identification of Man-made Regions in Unmanned Aerial Vehicle Imagery," *Proceedings of the 5th Automatic Target Recognizer System and Technology Symposium*, 1, 318–341.
Solka, J. L., Marchette, D. J., Rogers, G. W., Durling, E. C., Green, J. E., and Talsma, D. (1997b), "Region of Interest Identification in Unmanned Aerial Vehicle Imagery," *Emerging Applications of Computer Vision: SPIE*, 2962, 180–191.
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.
Worsley, K. J. (1994), "Local Maxima and the Expected Euler Characteristic of Excursion Sets of χ^2, F and t Fields," *Advances in Applied Probability*, 26, 13–42.
Young, T. Y., and Coraluppi, G. (1970), "Stochastic Estimation of a Mixture of Normal Density Functions Using an Information Criterion," *IEEE Transactions on Information Theory*, 16, 258–263.