ESTIMATION OF MIXTURE DENSITIES FROM HISTOGRAMS

David M. Rouse and H. Joel Trussell

Department of Electrical and Computer Engineering
North Carolina State University
Raleigh, NC 27695-7911

ABSTRACT

Many signals and statistical distributions are a mixture of component signals or distributions. Current methods for estimating the proportion of each component assume a parametric form for the components. We introduce nonparametric methods, based on projections onto convex sets, to address the many practical cases where parametric models are not applicable. Comparisons are made with parametric methods and discussed for special cases where both methods can be used.

1. INTRODUCTION

A recorded signal is often formed by a mixture of many signals from several classifiable sources. It is of interest in various applications to determine the proportion of the recorded signal that belongs to each class. For instance, spatial resolution limitations in remote sensing images inevitably possess pixels comprising various proportions of several spectral classes of ground cover [1]. A target often appears within a single pixel, mixed with its surroundings. One may detect the target's presence, or proportion, in that pixel using estimates of the target's spectral distribution. Another example of mixed signals occurs in network traffic across a particular node. Several internet applications exhibit characteristic distributions of packet sizes [2]. The total distribution of packet sizes across a network node is the mixture of the distributions of individual applications. With an accurate estimate of the proportion of each application, routers could give precedence to applications with time sensitive packets such as streaming media to improve the quality of service (QoS). Unusual network traffic at a node could indicate a potential security breach.

Previous research estimates class proportions of linearly mixed signals using the methods of maximum likelihood (ML) and total least squares (TLS) [3], [4]. However, the estimation of mixture densities currently assumes the mixture components form classical parametric distributions (i.e. Gaussian, Poisson, etc.), and the parameters of these distributions are estimated. Practical applications involving mixture densities rarely have components with classical parametric distributions. Nonparametric probability distributions of mixture components may be approximated by histograms formed from sampled data obtained for each component. These approximations can be used to estimate the proportion of each component in the recorded mixture if the uncertainty in the approximation is considered. This paper introduces the method of mixture density estimation from histograms using a set theoretic approach discussed in [5].

This paper presents a mathematical description of the problem. The estimation method of total least squares (TLS) is reviewed. The method projection onto convex sets (POCS) is reviewed and formed with a noise set derived from TLS. Results from experimental data comparing the EM Algorithm with the new method using POCS are provided.

2. DEFINITION OF THE PROBLEM

The problem of interest is to estimate the proportion of each of the K classes present. To do this, it is necessary to know the statistical distributions of the various classes. Since the distributions are nonparametric, these must be estimated from samples of the classes. We assume that we have K well defined classes and N_k samples of M-vectors from the k^{th} class $\{\mathbf{u}_{i,k}\}_{i=1}^{N_k}$. Note that each class may have a different number of samples and the total number is $N = \sum_{k=1}^{K} N_k$ samples. This can be thought of as a training set and the distribution of each class is estimated from these samples. Define estimates of the mean of each class as $\bar{\mathbf{s}}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} \mathbf{u}_{i,k}$.

Now consider a mixture of samples from the various classes. Let $\mathbf{s}_{i,k}$ be a vector of length M and a member of the k^{th} class. The linear mixture is the M-vector defined by

$$\mathbf{r} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N_k} \mathbf{s}_{i,k}. \tag{1}$$

We know that

$$E\{\mathbf{r}\} = \bar{\mathbf{S}}\mathbf{a}$$

where $\bar{\mathbf{S}} = \left[E\{\mathbf{s}_1\},\, E\{\mathbf{s}_2\},\, \cdots,\, E\{\mathbf{s}_k\}\right]$, $\mathbf{s}_k \in C_k$ and $\mathbf{a} = \left[N_1/N,\, N_2/N,\, \cdots, N_k/N\right]^T$. The obvious estimation of $\mathbf{a}$ uses the mean of the distributions combined with common maximum likelihood (ML) or minimum mean square error (MMSE) estimators. However, the true contribution from class k, $\frac{1}{N}\sum_{i=1}^{N_k}\mathbf{s}_{i,k}$, is only approximated by $\left(N_k/N\right)\bar{\mathbf{s}}_k$. This approximation is accounted for by inserting a perturbation term $\Delta\mathbf{s}_k$ for the estimate of the mean of class k. The perturbation term $\Delta\mathbf{s}_k$ is given by

$$\Delta\mathbf{s}_k = \bar{\mathbf{s}}_k - \frac{1}{N_k}\sum_{i=1}^{N_k}\mathbf{s}_{i,k},$$

and the perturbation is assumed to be an unbiased white, noise random process with variance σ_ν^2. However, the true variance of the m^{th} element of the k^{th} class $\Delta\mathbf{s}_k$ is $\sigma_\nu^2(m,k)$ for $1 \leq m \leq M$ and $1 \leq k \leq K$. Considering the deviations in the estimates of the mean of each class, the linear mixture $\mathbf{r}$ is more accurately written

$$\mathbf{r} = \left(\bar{\mathbf{S}} + \Delta\mathbf{S}\right)\mathbf{a}, \tag{2}$$

where $\Delta\mathbf{S} = \left[\Delta\mathbf{s}_1,\ \Delta\mathbf{s}_2,\ \cdots,\ \Delta\mathbf{s}_k\right]$.

Estimation methods that can take the perturbation into account include total least squares (TLS) and projection onto convex sets (POCS). MMSE can be formulated to accommodate this perturbation, but a solution is difficult. In some cases, there may be an additional noise added to the mixture that can be attributed to measurement uncertainty. For this case we have the model

$$\mathbf{r} = (\bar{\mathbf{S}} + \Delta\mathbf{S})\mathbf{a} + \boldsymbol{\eta}, \tag{3}$$

where $\boldsymbol{\eta}$ is signal independent noise, usually assumed to be independent and identically distributed from a zero mean, white noise random process with variance σ_η^2. The methods that can be used here are the same as for eq. (2). Note we can use eq. (2) to write

$$\mathbf{r} = \bar{\mathbf{S}}\mathbf{a} + \Delta\mathbf{S}\mathbf{a} = \bar{\mathbf{S}}\mathbf{a} + \boldsymbol{\upsilon},$$

where the noise is combined in the term $\boldsymbol{\upsilon}$. This may permit approximations that allow signal independent noise.

The nature of this mixture problem imposes constraints on $\mathbf{a}_{K\times 1}$. Since $\mathbf{a}$ represents proportions of the K classes in the recorded mixture, the elements of $\mathbf{a}$ must sum to one and lie on the interval $\left[0,1\right]$. The vector $\mathbf{a}$ is constrained to the set S_a defined in eq. (4).

$$S_a = \left\{\ \mathbf{a} \in \Re^K\ \middle|\ \sum_{k=1}^{K} a_k = 1,\quad a_k \geq 0\ \right\} \tag{4}$$

When the recorded signal $\mathbf{r}$ is a discrete mixture distribution composed of discrete probability distributions contained in the set of vectors $\left\{\mathbf{s}_k\right\}_{k=1}^{K}$, the elements of $\mathbf{r}$ must sum to one. The vector $\mathbf{r}$ represents a M bin probability histogram.

The TLS method is reviewed. Set theoretic estimation is presented to solve the TLS problem described by eq. (3) satisfying the additional constraints in eq. (4).

3. ESTIMATION METHODS

3.1. Total Least Squares

The TLS method minimizes the presence of the perturbations, $\Delta\mathbf{S}$ and $\boldsymbol{\eta}$, to find a solution $\hat{\mathbf{a}}$ to eq. (3). This method requires that $\Delta\mathbf{S}$ and $\boldsymbol{\eta}$ have the same variance. A weighted TLS method has been proposed when the variances of the noise sources differ. A solution found with the method of TLS is presented in accordance with [7].

3.1.1. TLS Solution

The elements $\mathbf{r}$ and $\bar{\mathbf{S}}$ in eq. (3) are precisely known. Thus, a solution to eq. (3) is an approximate solution to

$$\mathbf{r} \approx \bar{\mathbf{S}}\mathbf{a} \Rightarrow \bar{\mathbf{S}}\mathbf{a} - \mathbf{r} \approx 0. \tag{5}$$

A solution to this approximation is found by considering the equation

$$\left[\bar{\mathbf{S}};\mathbf{r}\right]\left[\begin{array}{c}\mathbf{a}\\ \text{-1}\end{array}\right] = 0. \tag{6}$$

The solution $\left[\hat{\mathbf{a}}^T;-1\right]^T$ must be in $\mathrm{Ker}\left(\left[\bar{\mathbf{S}};\mathbf{r}\right]\right)$. $\mathrm{Ker}\left(\mathbf{A}\right)$ denotes the nullspace of matrix $\mathbf{A}$. Reduce the rank of $\left[\bar{\mathbf{S}};\mathbf{r}\right]$ to find a solution in $\mathrm{Ker}\left(\left[\bar{\mathbf{S}};\mathbf{r}\right]\right)$. Create $\left[\hat{\mathbf{S}};\hat{\mathbf{r}}\right] \approx \left[\bar{\mathbf{S}};\mathbf{r}\right]$ by perturbing $\left[\bar{\mathbf{S}};\mathbf{r}\right]$ as little as possible.

To do this, define the singular value decomposition (SVD) of $\left[\bar{\mathbf{S}};\mathbf{r}\right]$ as follows

$$\left[\bar{\mathbf{S}};\mathbf{r}\right] = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T, \tag{7}$$

where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices, and $\boldsymbol{\Lambda}$ is a diagonal matrix ordered by descending singular values $\{\sigma_i\}_{i=1}^{N+1}$. Given that $\sigma_{N+1} \neq 0$, the matrix $\left[\bar{\mathbf{S}};\mathbf{r}\right]$ has rank $N+1$. In this case, $\mathbf{r}$ clearly is not contained in the column space of $\bar{\mathbf{S}}$. Reduce the rank of matrix $\left[\bar{\mathbf{S}};\mathbf{r}\right]$ to N to find a solution $\hat{\mathbf{a}}$.

The Eckart-Young-Mirsky theorem provides a matrix approximation using SVD of eq. (7). Let

$$\hat{\boldsymbol{\Lambda}} = \mathrm{diag}\left(\sigma_1,\ \cdots,\ \sigma_N,\ 0\right)$$

and define $\left[\hat{\mathbf{S}};\hat{\mathbf{r}}\right] = \mathbf{U}\hat{\boldsymbol{\Lambda}}\mathbf{V}^T$, the theorem gives the TLS correction

$$\sigma_{N+1} = \min_{\mathrm{rank}\left(\left[\hat{\mathbf{S}};\hat{\mathbf{r}}\right]\right)=N}\left\|\left[\mathbf{S};\mathbf{r}\right] - \left[\hat{\mathbf{S}};\hat{\mathbf{r}}\right]\right\|_F. \tag{8}$$

Respectively denoting $\mathbf{u}_i$ and $\mathbf{v}_i$ as the i^{th} columns of $\mathbf{U}$ and $\mathbf{V}$, we have

$$\left[\mathbf{S};\mathbf{r}\right] - \left[\hat{\mathbf{S}};\hat{\mathbf{r}}\right] = \left[\Delta\hat{\mathbf{S}};\boldsymbol{\eta}\right] = \sigma_{N+1}\mathbf{u}_{N+1}\mathbf{v}_{N+1}^T. \tag{9}$$

The last column of $\mathbf{V}$, $\mathbf{v}_{N+1} \in \mathrm{Ker}\left(\left[\hat{\mathbf{S}};\hat{\mathbf{r}}\right]\right)$, contains the TLS solution. Scale $\mathbf{v}_{N+1}$ to make its last component -1, or

$$\begin{bmatrix} \hat{\mathbf{a}} \\ \text{-1} \end{bmatrix} = \frac{-1}{v_{N+1,N+1}} \mathbf{v}_{N+1}. \quad (10)$$

Thus, the TLS solution is

$$\hat{\mathbf{a}} = \frac{-1}{v_{N+1,N+1}} \left[v_{1,N+1},\, v_{2,N+1},\, \cdots,\, v_{N,N+1} \right]^T. \quad (11)$$

The minimum norm total least squares (MN-TLS) solution shown in [7] is determined from the following

$$\min_{\Delta\mathbf{E},\mathbf{a}} \|\Delta\mathbf{E}\|_F^2 \quad \textit{subject to} \quad \left[\mathbf{E} + \Delta\mathbf{E}\right] \begin{bmatrix} \mathbf{a} \\ \text{-1} \end{bmatrix} = 0, \quad (12)$$

where $\mathbf{E}_{M\times(N+1)} = \left[\bar{\mathbf{S}}; \mathbf{r}\right]$ and $\Delta\mathbf{E}_{M\times(N+1)} = \left[\Delta\mathbf{S}; \boldsymbol{\eta}\right]$.

3.2. Projection onto Convex Sets

Previous estimation methods satisfy some of the constraints known about the signal $\mathbf{a}$. However, the constraints specific to the problem of mixture densities are not addressed. A solution to eq. (3) including the constraints in eq. (4) is presented utilizing sets corresponding to the known characteristics of $\mathbf{a}$. The intersection of these sets provides a set of feasible solutions rather than a unique solution.

Set theoretic estimation utilizes *a priori* knowledge of a signal to generate a feasible estimate [6]. The solution is in the intersection of a collection of constraint sets. A solution may be obtained by the well-known method of sequential projections onto convex sets (POCS) [6]. We need only define the constraint sets used for this application.

3.2.1. Noise Variance Set

For the noise constraint, the set with regard to the ML residual definition is modified since $\mathbf{S} = \bar{\mathbf{S}} + \Delta\mathbf{S}$. Applying the triangle inequality and the additional noise parameter σ_v^2 modifies the ML residual set to

$$S_\eta = \left\{ \mathbf{a} \,\middle|\, \left\|\mathbf{r} - \bar{\mathbf{S}}\mathbf{a}\right\|^2 \le \sigma_\eta^2 + \sigma_v^2 \right\}, \quad (13)$$

where $\sigma_\eta^2 = E\left\|\boldsymbol{\eta}\right\|^2$ and $\sigma_v^2 = E\left\|\Delta\mathbf{S}\mathbf{a}\right\|^2$. Alternatively, as shown in [5] the set S_η may be defined based on weighted total least squares (TLS) as

$$S_{TLS} = \left\{ \mathbf{a} \,\middle|\, \exists\{\Delta\mathbf{S}, \boldsymbol{\eta}\} \ni \left(\bar{\mathbf{S}} + \Delta\mathbf{S}\right)\mathbf{a} = \mathbf{r} + \boldsymbol{\eta}, \right.$$
$$\left. \tau\left\|\Delta\mathbf{S}\right\|_F^2 + \left\|\boldsymbol{\eta}\right\|^2 \le \nu \right\}. \quad (14)$$

where the parameters τ and ν are determined by statistical properties of $\Delta\mathbf{S}$ and $\boldsymbol{\eta}$. It was shown in [5] that S_{TLS} may also be defined by

$$S_{TLS} = \left\{ \mathbf{a} \in \Re^N \,\middle|\, \left\|\bar{\mathbf{S}}\mathbf{a} - \mathbf{r}\right\|^2 - \frac{\nu}{\tau}\left\|\mathbf{a}\right\|^2 - \nu \le 0 \right\}, \quad (15)$$

where τ and ν are chosen to satisfy the following

$$\nu = \tau E\left\|\Delta\mathbf{S}\right\|_F^2 + E\left\|\eta\right\|^2 \qquad \tau \ge \frac{E\left\|\eta\right\|^2}{\sigma_N\left(\bar{\mathbf{S}}\right) - E\left\|\Delta\mathbf{S}\right\|_F^2}, \quad (16)$$

where $\sigma_N\left(\bar{\mathbf{S}}\right)$ denotes the smallest singular value of $\bar{\mathbf{S}}$.

3.2.2. Other Sets

The set S_a defined in eq. (4) is decomposed into two appropriate sets. The set S_Σ defines the summation to one constraint of the vector elements, and the set S_n defines the set of nonnegative vectors.

3.2.3. Projection onto Sets

The projection onto the set S_{TLS} is

$$\mathbf{a}_0 = \left[\mathbf{I} + \lambda_\eta\left(\bar{\mathbf{S}}^T\bar{\mathbf{S}} - \frac{\nu}{\tau}\mathbf{I}\right)\right]^{-1}\left(\hat{\mathbf{a}} + \lambda_\eta\bar{\mathbf{S}}^T\mathbf{r}\right), \quad (17)$$

where $\hat{\mathbf{a}} \notin S_{TLS}$ and λ_η is the Lagrange multiplier satisfying $f(\mathbf{a}_0)$, where [5] defines $f(\cdot)$ as

$$f(\mathbf{a}) \equiv \left\|\bar{\mathbf{S}}\mathbf{a} - \mathbf{r}\right\|^2 - \frac{\nu}{\tau}\left\|\mathbf{a}\right\|^2 - \nu.$$

The projection onto the set S_Σ is

$$\mathbf{a}_0 = \hat{\mathbf{a}} - \frac{1}{K}\left(\hat{\mathbf{a}}^T\mathbf{1} - 1\right)\mathbf{1}_{K\times 1}, \quad (18)$$

where $\hat{a} \notin S_\Sigma$.

The projection onto the nonnegativity set S_n is performed by replacing negative elements of $\hat{\mathbf{a}}$ with zeros to form $\mathbf{a}_0$.

4. SIMULATIONS

The performance of the set theoretic approach to estimate the proportion of the classes present in a mixture was tested for mixtures containing parametric and nonparametric distributions.

For parametric distributions, the expectation-minimization (EM) algorithm provides a model to compare results from the proposed set theoretic approach. A classical parametric mixture distribution was formed for simulation using 1000 realizations of $K = 2$, Gaussian distributions with unique means and variances. For the POCS method, each class estimate was generated from 10 sets of histograms formed using 1000 realizations from each Gaussian distribution known to exist in the mixture.

A nonparametric mixture density was formed using a combination of $K = 2$ nonstandard distributions. Estimates for each class were generated from 10 sets of histograms formed using 100 realizations from each nonstandard distribution known to exist in the mixture.

Both simulations implemented the following for the POCS estimates. The mean class distributions were set as column vectors of length $M = 100$. Noise power of -50dB and -60dB approximated σ_v^2 and σ_η^2, respectively. The parameter τ was set at 1000 times the value defined in eq. (16). Mean-squared error (MSE) statistics were obtained for each method.

5. RESULTS

Estimation results for mixing proportions in parametric and nonparametric distributions listed in Table 1 indicate the precision of the estimates of $\mathbf{a}$. The method of POCS using the modified ML, from eq. (13), and TLS, from eq. (15), residual sets was compared to the EM-algorithm found in [4] for parametric distributions. Note that the EM algorithm estimates the mixture proportions from the unbinned recorded data, but the POCS methods estimate the mixture proportions from a normalized histogram of the recorded data. This should be considered when evaluating the error measurements. Figure 1 shows estimation results for nonparametric distributions and the corresponding means of the class distributions in the mixture density. The POCS-TLS errors for the nonparametric mixture distribution in Figure 1 correspond to those listed in Table 1.

Table 1. MSE of $\mathbf{a}$ (dB)

	Parametric Distributions
EM-Algorithm	-50.18
POCS-ML	-32.51
POCS-TLS	-33.27

	Nonparametric Distributions
POCS-ML	-33.28
POCS-TLS	-35.37

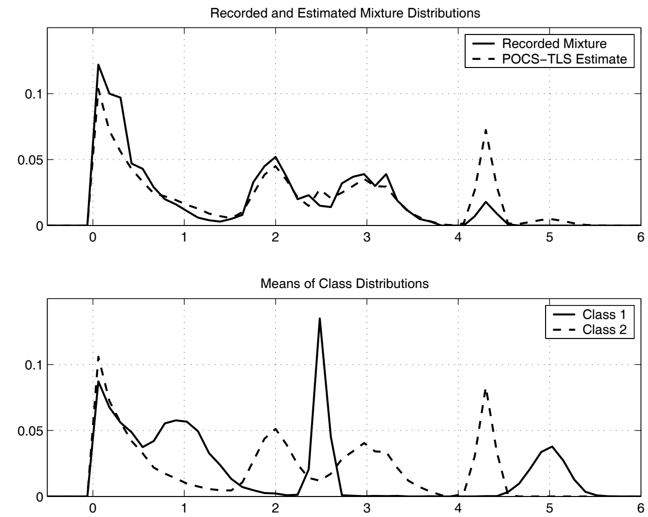

Fig. 1. Nonparametric Results

6. CONCLUSIONS

Our nonparametric methods based on POCS compared well to parametric methods The EM estimation is expected to perform better for parametric distributions, since it estimates the parameters of the distributions and the mixture proportions. This comparison should be taken more as an indication of the viability of the new method, since the assumptions on the two methods are significantly different. The parametric methods assumes knowledge of the form of the underlying distributions. The nonparametric methods assume estimates of the means of the distributions. The quality of these estimates will significantly influence their performance. The quality will also depend on the variation on the sample distributions for each realization. When the means are accurate, the nonparametric methods can better the EM estimation. As the estimates of the means become less accurate the estimates of the mixture proportions degrade as expected. However, for the nonparametric case, the parametric methods cannot be used, leaving the new method alone in the field.

7. REFERENCES

[1] *IEEE Signal Processing Magazine, Special Issue on Hyperspectral Imaging*, vol. 19, Jan. 2002.

[2] C. Trivedi, H. J. Trussell, A. Nilsson, and M. Chow, “Implicit traffic classification for service differentiation,”ITC Specialist seminar, July 2002, Wurzburg, Germany.

[3] Y. H. Hu, H. B. Lee, and F. L. Scarpace, “Optimal linear spectral unmixing,” *IEEE Tran. on Geoscience and Remote Sensing*, vol. 37, pp. 639–644, 1999.

[4] G. J. McLachlan and D. Peel, *Finite Mixture Models*. New York: Wiley, 2000.

[5] G. Sharma and H. J. Trussell, “Set theoretic signal restoration using and error in variables criterion,” *IEEE Tran. on Image Proc.*, vol. 6, pp. 1692–1697, 1997.

[6] P. L. Combettes, “The Foundations of Set Theoretic Estimation,” *Proc. IEEE*, vol. 81, pp. 182-208, Feb. 1993

[7] S. van Huffel and J. Vandewalle, *The Total Least Squares Problem*. Philadelphia, PA: SIAM, 1991.