BAYESIAN FRAMEWORK FOR UNSUPERVISED CLASSIFICATION WITH APPLICATION TO TARGET TRACKING

R. L. Kashyap and Srinivas Sista

School of Electrical & Computer Engineering
Purdue University, West Lafayette, IN 47907-1285
{kashyap,sista}@ecn.purdue.edu

ABSTRACT

We have given a solution to the problem of unsupervised classification of multidimensional data. Our approach is based on Bayesian estimation which regards the number of classes, the data partition and the parameter vectors that describe the density of classes as unknowns. We compute their MAP estimates simultaneously by maximizing their joint posterior probability density given the data. The concept of partition as a variable to be estimated is a unique feature of our method. This formulation also solves the problem of validating clusters obtained from various methods. Our method can also incorporate any additional information about a class while assigning its probability density. It can also utilize any available training samples that arise from different classes.

We provide a descent algorithm that starts with an arbitrary partition of the data and iteratively computes the MAP estimates. The proposed method is applied to target tracking data. The results obtained demonstrate the power of Bayesian approach for unsupervised classification.

1. INTRODUCTION

In unsupervised classification, the given data $Z = \{\boldsymbol{z_i}, i = 1, \cdots, N\}, \boldsymbol{z_i} \in R^m$ has to be partitioned into mutually exclusive and totally inclusive subsets of Z namely $\boldsymbol{c} = \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_s\}, \boldsymbol{c}_k \subseteq Z$ so that all the members belonging to a class are close to each other in some sense. The choice of s is itself a problem. The solution should include a compact description of each class so that a new unlabeled data point can be classified easily. Next the methodology should include the validation of the partition, i.e does the given partition adequately explain the data? Given two different partitions, which one of them gives a better explanation of the data?

This work has been partially supported by National Science Foundation under contract IRI 9619812 and the office of Naval Research under contract N00014-91-J-4126.

When s is fixed, partitions are given by the so called clustering algorithms [1, 2, 3, 4]; however there is no guarantee that the partition really explains the data. The criterion function used in the algorithms focus only on the centroid of clusters, not on the shape and orientation of the clusters. Further, the criterion function usually has numerous local minima and many methods stop with obtaining one arbitrary local minimum. Currently there is no method to compare different cluster sets derived for the same data obtained from different methods.

In the Bayes approach given in this paper the partition $\boldsymbol{c}$ is itself regarded as a variable to be chosen from the appropriate space. When s is known, $\boldsymbol{c} = \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_s\}, \boldsymbol{c} \in \Omega_{s,s}$, the set of all partitions of set Z so that none of the sets $\boldsymbol{c}_k$ in $\boldsymbol{c}$ are null. When s is not specified, $s \leq s_0$, then $\boldsymbol{c} = \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_{s_0}\}, \boldsymbol{c} \in \Omega_{s_0}$, the set of all partitions of Z into s_0 subsets. The members of class k are described by the probability density $p_k(\boldsymbol{z_i} \mid \boldsymbol{\theta}_k)$, p_k is a known function and $\boldsymbol{\theta}_k$ is a vector parameter whose values have to be determined, $\boldsymbol{\theta}_k \in R^{n_k}$. The unknowns are $\{\boldsymbol{c} = \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_s\}, \boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_s\}\}$.

The Bayes approach allows us to estimate s, the number of classes given that $s \leq s_0$. Correspondingly the best partition $\boldsymbol{c}$ has to be searched in the space $\boldsymbol{c} \in \Omega_{s_0}$. It also solves the problem of partition comparison or cluster validation. Two different partitions $\boldsymbol{c}$ and $\boldsymbol{c}'$ involving different values of s can be compared by computing the ratio of the corresponding posterior probabilities $P(\boldsymbol{c} \mid Z)$ and $P(\boldsymbol{c}' \mid Z)$. Our method can also utilize any additional information on the classes in assigning the probability density function p_k. For example, when all the members $\boldsymbol{z_i}$ are clustered tightly around a straight line or a convex curve or a 2-D plane.

2. OPTIMAL PARTITION WITH A GIVEN NUMBER OF CLASSES

Let the data set be $Z = \{\boldsymbol{z_i}, i = 1, \cdots, N\}, \boldsymbol{z_i} \in R^m$ whose members are statistically independent. Let s be

the number of distinct classes in Z, s is known to us. Let the s associated probability densities be $p_k(\boldsymbol{z_i} \mid \boldsymbol{\theta}_k), \boldsymbol{\theta}_k \in R^{n_k}, k = 1, \cdots, s$. Let the set $\boldsymbol{c} = \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_s\}$ be a partition of Z into s classes such that

$$\begin{array}{c} \boldsymbol{c}_k \subseteq Z, \ \forall k = 1, \cdots, s; \ \boldsymbol{c}_i \cap \boldsymbol{c}_j = \text{Null}, i \neq j \\ \bigcup_{k=1}^{s} \boldsymbol{c}_k = Z. \quad \boldsymbol{c}_k \neq \text{ Null } \forall k = 1, \cdots, s. \end{array} \tag{1}$$

Each $\boldsymbol{c}_k$ is a subset of Z whose members are described by the density p_k. Let $\Omega_{s,s}$ be the set of all possible distinct partitions of Z obeying (1). $\boldsymbol{c}$ and $\boldsymbol{c}'$ are different partitions if they are different sets. The number of distinct partitions, which is the cardinality of $\Omega_{s,s}$ is

$$\#\Omega_{s,s} = \frac{1}{s!}\sum_{i=0}^{s}(-1)^i \binom{s}{i} (s-i)^N \approx \frac{s^N}{s!} \tag{2}$$

$\boldsymbol{c}$ and $\boldsymbol{\theta}_k, k = 1, \cdots, s$ are the variables to be estimated. We regard $\boldsymbol{c} \in \Omega_{s,s}, \boldsymbol{\theta}_k \in R^{n_k}, k = 1, \cdots, s$ as independent random variables. $P(\boldsymbol{c})$, the prior probability associated with $\boldsymbol{c}$ is same for all $\boldsymbol{c}$; $P(\boldsymbol{c}) = \frac{1}{\#\Omega_{s,s}}, \ \ \forall \boldsymbol{c} \in \Omega_{s,s}$. Let $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_s\}$. Let $p(\boldsymbol{\theta}_k)$ be the prior probability density of $\boldsymbol{\theta}_k$ such that each component is uniformly distributed. Since the priors of $\boldsymbol{\theta}$ and $\boldsymbol{c}$ are uniform, the MAP estimates $(\boldsymbol{c}^*, \boldsymbol{\theta}^*)$ are given by

$$(\boldsymbol{c}^*, \boldsymbol{\theta}^*) = \text{Arg} \max_{\boldsymbol{c},\boldsymbol{\theta}} P(Z \mid \boldsymbol{c}, \boldsymbol{\theta}). \tag{3}$$

Since the data Z is independent, the joint density of Z has the following form:

$$P(Z \mid \boldsymbol{c}, \boldsymbol{\theta}) = \prod_{k=1}^{s} \left(\prod_{\boldsymbol{z_i} \in \boldsymbol{c}_k} p_k(\boldsymbol{z_i} \mid \boldsymbol{\theta}_k). \right) \tag{4}$$

Let $f_k(\boldsymbol{z_i}, \boldsymbol{\theta}_k) = -2 \ln p_k(\boldsymbol{z_i} \mid \boldsymbol{\theta}_k)$, then

$$J(\boldsymbol{c}, \boldsymbol{\theta}) = -2 \ln P(Z \mid \boldsymbol{c}, \boldsymbol{\theta}) = \sum_{k=1}^{s} \sum_{\boldsymbol{z_i} \in \boldsymbol{c}_k} f_k(\boldsymbol{z_i}, \boldsymbol{\theta}_k). \tag{5}$$

and

$$(\boldsymbol{c}^*, \boldsymbol{\theta}^*) = \text{Arg} \min_{\boldsymbol{c},\boldsymbol{\theta}} J(\boldsymbol{c}, \boldsymbol{\theta}) \tag{6}$$

$J(\boldsymbol{c}, \boldsymbol{\theta})$ has interesting extremal properties.

For a fixed $\boldsymbol{\theta}$ the value of $\boldsymbol{c}$ which minimizes $J(\boldsymbol{c}, \boldsymbol{\theta})$ w.r.t $\boldsymbol{c}$ can be obtained using

$$\begin{array}{rcl} \hat{\boldsymbol{c}}_{\boldsymbol{\theta},k} & = & \{\boldsymbol{z_i} : f_k(\boldsymbol{z_i}, \boldsymbol{\theta}_k) \leq f_u(\boldsymbol{z_i}, \boldsymbol{\theta}_u), \\ & & \forall k \neq u, u = 1, \cdots, s\}, k = 1, \cdots, s \end{array} \tag{7}$$

Similarly for a fixed $\boldsymbol{c}$, the minimizing value of $\boldsymbol{\theta}$ is unique and it can be obtained using

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{c},k} = \min_{\boldsymbol{\theta}_k \in R^{n_k}} \sum_{\boldsymbol{z_i} \in \boldsymbol{c}_k} f_k(\boldsymbol{z_i}, \boldsymbol{\theta}_k), \quad k = 1, \cdots, s. \tag{8}$$

When p_k are given by $p_k(\boldsymbol{z_i} \mid \boldsymbol{\theta}_k) \sim \text{Gauss}(\boldsymbol{\phi}_k, \boldsymbol{r}_k)$, $\boldsymbol{\theta}_k = \{\boldsymbol{\phi}_k, \boldsymbol{r}_k\}$ an explicit expression for $\hat{\boldsymbol{\theta}}_{\boldsymbol{c},k}$ can be given because of the structure of f_k as follows

$$\begin{array}{l} f_k(\boldsymbol{z_i}, \boldsymbol{\theta}_k) = \\ (\boldsymbol{z_i} - \boldsymbol{\phi}_k)^T \boldsymbol{r}_k^{-1} (\boldsymbol{z_i} - \boldsymbol{\phi}_k) + \ln \mid \det \boldsymbol{r}_k \mid + m \ln 2\pi \end{array} \tag{9}$$

Let $\boldsymbol{\theta}_k = (\boldsymbol{\phi}_k, \boldsymbol{r}_k)$, where $\boldsymbol{\phi}_k$ is an m-vector and $\boldsymbol{r}_k$ is an $m \times m$ covariance matrix. Then

$$\begin{array}{rcl} \hat{\boldsymbol{\phi}}_{\boldsymbol{c},k} & = & \dfrac{1}{N_{1k}} \displaystyle\sum_{\boldsymbol{z_i} \in \boldsymbol{c}_k} \boldsymbol{z_i} \\ \hat{\boldsymbol{r}}_{\boldsymbol{c},k} & = & \dfrac{1}{N_{1k}} \displaystyle\sum_{\boldsymbol{z_i} \in \boldsymbol{c}_k} (\boldsymbol{z_i} - \hat{\boldsymbol{\phi}}_{\boldsymbol{c},k})(\boldsymbol{z_i} - \hat{\boldsymbol{\phi}}_{\boldsymbol{c},k})^T \\ N_{1k} & = & \#\boldsymbol{c}_k \end{array} \tag{10}$$

A simple descent algorithm is given for finding a local minimum of $J(\boldsymbol{c}, \boldsymbol{\theta})$. It is done by changing $\boldsymbol{\theta}$ and $\boldsymbol{c}$ alternatively using expressions (7) and (8), each time having a reduction in $J(\boldsymbol{c}, \boldsymbol{\theta})$. Note that a local minimum need not be a global minimum, since we perturb only $\boldsymbol{c}$ or $\boldsymbol{\theta}$ at one time, not simultaneously.

Since the determination of $\boldsymbol{\theta}_k$ utilizing all $\boldsymbol{z_i}$ in $\boldsymbol{c}_k$ involves the inversion of a matrix, assume that the number of members in $\boldsymbol{c}_k$ must be greater than $2n_k$. Let us call this assumption (A1).

Descent Algorithm(with assumption A1)

1. Let $\boldsymbol{c}^j = (\boldsymbol{c}_1^j, \cdots, \boldsymbol{c}_s^j)$ and $\boldsymbol{\theta}^j = (\boldsymbol{\theta}_1^j, \cdots, \boldsymbol{\theta}_s^j)$ be estimates at the end of j^{th} iteration. Choose $\boldsymbol{c}^1$ arbitrarily, perhaps from a solution of a clustering algorithm with random seeds.
2. Given $\boldsymbol{c}^{(j)}$, compute $\boldsymbol{\theta}^{(j)}$ using the formula in (8).
3. Given $\boldsymbol{\theta}^{(j)}$, compute $\boldsymbol{c}^{(j+1)}$ using (7).
4. Stop if $\boldsymbol{c}^{(j)} = \boldsymbol{c}^{(j+1)}$; otherwise goto 2.

End.

Note that the computational effort for finding a local minimum is very little. It involves the inversion of a matrix of relatively small dimension in (8) and data comparisons in (7). The proof of convergence of the descent algorithm is given in [5].

3. CHOICE OF S, THE NUMBER OF DISTINCT CLASSES

The problem of choosing the value of s is known as model order identification or cluster validation. A popular method is to use the Akaike's information criterion. However, it has been shown in [6] that this criterion does not yield consistent estimates. In our method we obtain

the estimate of s via Bayesian estimation by considering s also as a random variable.

Using the Bayes formalism we will compare all the partitions $\boldsymbol{c}$ of Z in $\Omega_{s,s}$ for s, $1 \leq s \leq s_0$, s_0 being known and find the best partition, and incidentally the best value of s. So s is included as an unknown to be estimated, $1 \leq s \leq s_0$. The optimal Bayes estimator of $(s, \boldsymbol{c}, \boldsymbol{\theta})$ is given by

$$(s^*, \boldsymbol{c}^*_{s^*}, \boldsymbol{\theta}^*) = \text{Arg} \min_{1 \leq s \leq s_0} \left\{ \min_{\boldsymbol{c} \in \Omega_{s,s}} \min_{\boldsymbol{\theta}_k \in R^{n_k}} H_s \right\} \quad (11)$$

where

$$H_s = -\ln \left\{ p(Z \mid s, \boldsymbol{c}, \boldsymbol{\theta}) P(\boldsymbol{c} \mid s) \left(\prod_{k=1}^{s} p(\boldsymbol{\theta}_k \mid s) \right) P(s) \right\}$$

The prior probabilities of s and $\boldsymbol{c}$ given s are chosen as follows:

$$P(s) = 1/s_0, \ s = 1, \cdots, s_0 \quad (12)$$

$$P(\boldsymbol{c} \mid s) = \frac{1}{\#\Omega_{s,s}}, \quad \sum_{\boldsymbol{c} \in \Omega_{s,s}} P(\boldsymbol{c} \mid s) = 1 \quad (13)$$

And the prior probability of each component in $\boldsymbol{\theta}_k$ is uniform and equals $1/L_k$. Since L_k is the prior density of $\boldsymbol{\theta}_k$, it should cover the total range of all the components of $\boldsymbol{\theta}_k$.

Comparing partitions with different values of s

Suppose we have 2 partitions $\boldsymbol{c}^1 \in \Omega_{s_1,s_1}$ and $\boldsymbol{c}^2 \in \Omega_{s_2,s_2}$ with the number of classes s_1 and s_2 respectively. We can compare the probabilities $P(s_k, \boldsymbol{c}^k, \boldsymbol{\theta}^{*k} \mid Z), k = 1, 2$ to decide which partition is better. We compute the log likelihood ratio $\ln \left(\frac{P(s_1, \boldsymbol{c}^1, \boldsymbol{\theta}^{*1} | Z)}{P(s_2, \boldsymbol{c}^2, \boldsymbol{\theta}^{*2} | Z)} \right)$.

4. EXPERIMENTAL RESULTS

Multi-sensor target tracking: Multiple sensors send observations $\boldsymbol{z_i} = (y_i, x_i), i = 1, \cdots, N$ to the central station [7, 8]. There could be multiple targets in the atmosphere and their number could be variable at any given time. The raw data of 120 points is shown in Figure 1(a). The x-coordinate is related to time. The figure shows all the observations collected up to a time t_1. We have not shown time in the graph. As time progresses there is more data. There is no target label attached to each observation. It is known apriori that the trajectory of a target obeys some parametric curve in the $X - Y$ plane; straight line, parabola etc. For simplicity we consider a straight line. There are also observations caused purely by noise, the clutter. Note that one trajectory is completely inside the clutter. Moreover the range of this trajectory is much less than that of others. The problem is to identify the number of targets, their tracks and the clutter points. Intersection of the trajectories in the figure indicates intersection in feature space, not in real time.

Each trajectory is parametrized by a line $L(\beta, \gamma, \rho)$ and obeys the equation

$$y_i = \beta x_i + \gamma + \text{Gauss}(0, \rho), \quad i = 1, \cdots, N \quad (14)$$

x_i are uniformly distributed in the range $[0, 10]$. The three line trajectories are $L_1(0.4, 5, 0.01), L_2(-0.3, 9, 0.01)$, $L_3(0.1, 2.1, 0.0025)$. The clutter is modeled by a Gaussian distribution given by $\text{Gauss}\left(\begin{bmatrix} 5 \\ 4 \end{bmatrix}, \begin{bmatrix} 0.4 & 0.2 \\ 0.2 & 1.2 \end{bmatrix} \right)$. There are 30 points in each trajectory class as well as the clutter class, a total of 120 points.

Results with fuzzy clustering

$s = 4$: The result with $s = 4$ is given in Figure 1(e). The clustering captures only one of the three line trajectories. One cluster combines parts of the 2 lines of the data, before the intersection. The other cluster captures the other two halves of the line clusters in the data. This usually happens with most clustering algorithms because they do not use the available information that the trajectories are straight lines or parabolas etc.

$s = 5$: The result with $s = 5$ is given in Figure 1(f). This clustering is also erroneous. It doesn't identify any line trajectories correctly. The clusters corresponding to the clutter and the trajectory within it are subdivided into two clusters without the trajectory being identified.

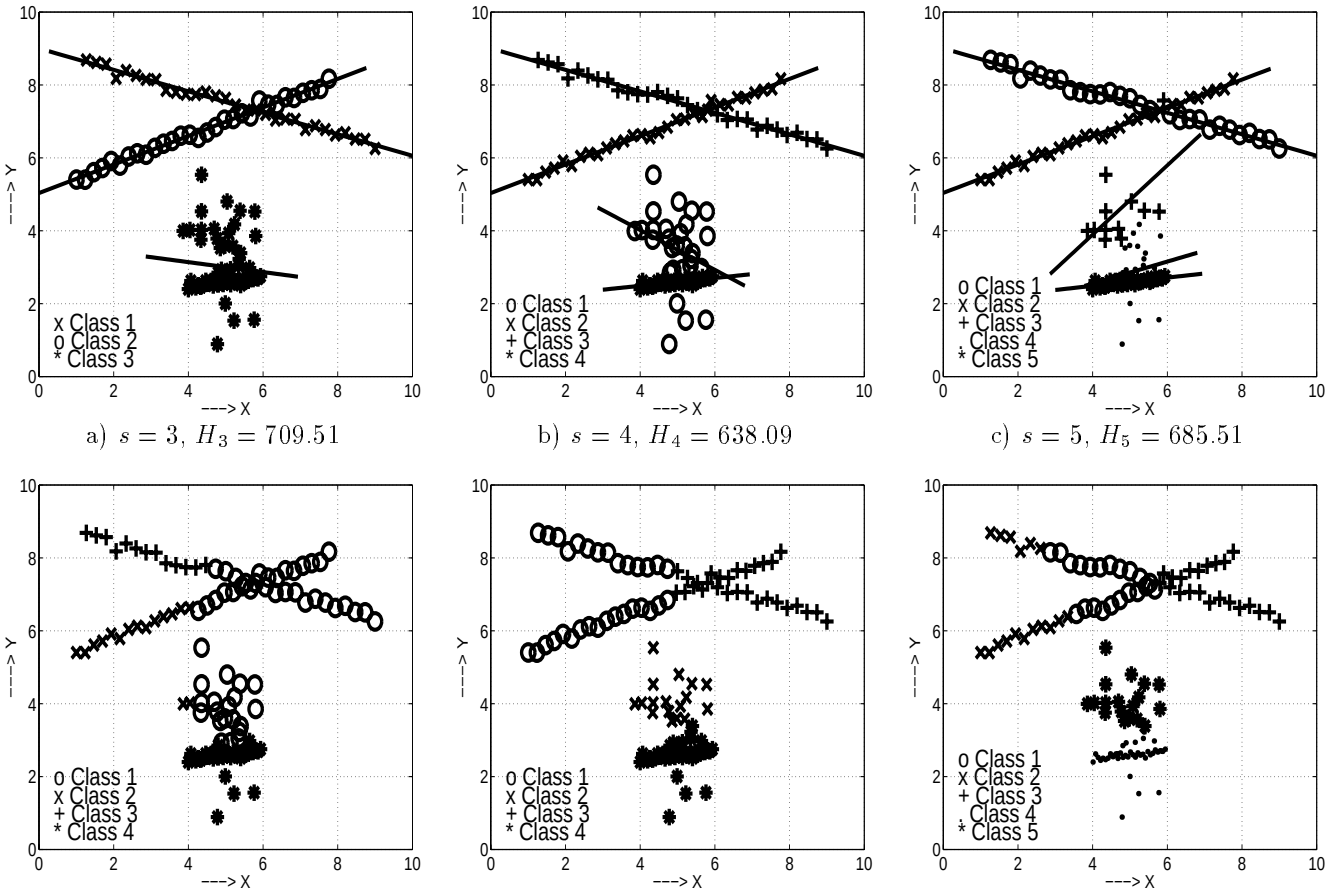
Results with Bayesian method

$s = 4$: All density families p_k are multivariate densities. The best local minimum, shown in Figure 1(b), has $H_s = 638.09$. Note that our method captures the four classes correctly. Even the trajectory within the clutter is identified correctly.

$s = 5$: The result associated with best local minimum is in Figure 1(c). Note the result divides the data of smaller trajectory and the clutter into 3 clusters, correctly finding the clusters of two big lines. $H_s = 685.51$. Notice that H_5, the H-statistic with $s = 5$ is much larger than H_4 indicating that the correct value of s is 4.

5. CONCLUSION

We proposed a solution to the problem of unsupervised classification of multidimensional data based on Bayesian estimation. The new feature of our method is, we regard the data partition as a variable to be estimated. We developed a Bayesian framework to estimate the number of classes, the class parameters and the data partition simultaneously. The cluster validation problem was formally addressed. We presented an example with target



a) $s = 3$, $H_3 = 709.51$ b) $s = 4$, $H_4 = 638.09$ c) $s = 5$, $H_5 = 685.51$

d) Initial for $s = 4$ e) Clustering, $s = 4$ f) Clustering, $s = 5$

Figure 1: Results of applying the Bayesian and fuzzy clustering methods to target tracking data. Data size $N = 120$ coming from 4 classes. (a,b,c) Classifications corresponding to best local minima obtained using Bayesian method with number of distinct classes $s = 3, 4$ and 5 respectively. (b) Classification with the lowest minimum amongst best minima with various values of s. (d) Initial random partition with $s = 4$ that converges to the result in (b). (e,f) Classifications obtained using Fuzzy clustering with number of classes $s = 4$ and 5 respectively.

tracking data. Future work will focus on robust regression, which is a special case of unsupervised classification with two classes namely inliers and outliers as well as image and video segmentation.

6. REFERENCES

[1] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.

[2] E. Ruspini, "Numerical Methods for Fuzzy Clustering," *Inf. Sci.*, Vol. 2, pp. 319–350, 1970.

[3] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms.* New York: Plenum Press, 1981.

[4] Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Comm.*, COM-28, pp. 84–95, January 1980.

[5] R. L. Kashyap and Srinivas Sista, "Unsupervised Classification and Choice of Classes: Bayesian Approach," *Technical Report TR-ECE 98-12*, School of Electrical and Computer Engineering, Purdue University, July 1998.

[6] R. L. Kashyap, "Inconsistency of the AIC rule for estimating the Order of Autoregressive Moving Average Models," *IEEE Trans. Automat. Contr.*, Vol. AC-25, No. 5, October 1980.

[7] A. Satish and R. L. Kashyap, "Estimation of Singularities for Intercept Point Forecasting," *IEEE Trans. on Aerospace and Electronic Systems*, Vol. 32, No. 4, pp. 1301–1309, October 1996.

[8] K. Wang, "An assessment of tactical surface-to-air missile midcourse guidance technology," In *Proceedings of the 1991 American Control Conference*, Vol. 1, pp. 854–855, 1991.