Sensor Integration for Classification

Steven Kay and Quan Ding
Department of Electrical, Computer,
and Biomedical Engineering
University of Rhode Island, U.S.A.
Email: kay@ele.uri.edu, dingqqq@ele.uri.edu

Muralidhar Rangaswamy
Air Force Research Laboratory Sensors Directorate
Wright-Patterson Air Force Base, U.S.A.
Email: Muralidhar.Rangaswamy@wpafb.af.mil

***Abstract*—In the problem of sensor integration, an important issue is to estimate the joint PDF of the measurements of sensors. However in practice, we may not have enough training data to have a good estimate. In this paper, we have constructed the joint PDF using an exponential family for classification. This method only requires the PDF under a reference hypothesis. Its performance has shown to be as good as the estimated maximum a posteriori probability classifier which requires more information. This shows a wide application of our method in classification because less information is needed than existing methods.**

***Index Terms*—Exponential family, classification, joint PDF, sensor integration.**

I. Introduction

Distributed detection/classification systems have been widely used in many applications such as radar, sonar, wireless sensor networks, and medical diagnosis. Since multiple sensors will collect more information than a single sensor does, a better decision is expected to be made. In classification, it is well known that the maximum a posteriori probability (MAP) classifier minimizes the probability of error [1]. However, the MAP rule requires the complete knowledge of the joint probability density functions (PDFs) of the measurements from sensors under each hypothesis, which in practice may not be available. Hence, it is important in sensor integration to find appropriate estimates of the joint PDFs under each hypothesis, and the estimates should contain all the available information.

In many works, people assume that the marginal PDFs of the measurements from each sensor are known. One commonly used method is to simply assume that the measurements are independent, and the joint PDF is just the product of the marginal PDFs [2], [3]. This is equivalent to the product rule in combining classifiers, and it is a severe rule as shown in [4]. Another concern is that the correlation among the measurements is neglected by assuming independence. So some approaches that consider the dependence among the measurements have been proposed. A copula based method that estimates the joint PDF from the marginal PDFs is used in [5], [6]. The exponentially embedded families (EEFs) that asymptotically minimize the Kullback-Leibler (KL) divergence between the true PDF and the estimated PDF is proposed in [7].

Note that the marginal PDFs are required in the above mentioned approaches. However, we may not even have enough training data in practice to have an accurate estimate of the marginal PDFs, especially when the sensor outputs have high dimensions. In this paper, we construct the joint PDF using an exponential family. The construction only requires a reference PDF and it incorporates all the available information. It can be shown that the constructed PDF is asymptotically the optimal one in the sense that it is asymptotically closest to the true PDF in KL divergence.

By maximizing the constructed PDF over the signal parameters, our classifier can be easily derived. The performance of our method is compared to that of the estimated MAP classifier, which assumes that the true joint PDF is known except for the unknown parameters. We present an example in which their performances appear to be the same. Note that our method assumes less information than the estimated MAP classifier does. This shows that our method has many applications for distributed systems in practice.

The paper is organized as follows. In Section 2, we introduce a distributed classification problem. In Section 3, we construct the joint PDF by an exponential family and apply it to the classification problem. An example is given in Section 4. In Section 5, the performances of our method and the estimated MAP classifier are compared via simulation. Conclusions are drawn in Section 6.

II. Problem Statement

Consider the classification problem where we have two distributed sensors whose outputs $\mathbf{T}_1(\mathbf{x})$ and $\mathbf{T}_2(\mathbf{x})$ are transformations of the underlying samples $\mathbf{x}$ that are unobservable. We need to decide from among M candidate hypotheses $\mathcal{H}_i$ for $i = 1, 2, \ldots, M$. Assume that there is a reference hypothesis $\mathcal{H}_0$ (usually it is the hypothesis with noise only) and we have enough training data $\mathbf{T}_{1_n}(\mathbf{x})$'s and $\mathbf{T}_{2_n}(\mathbf{x})$'s under $\mathcal{H}_0$ to accurately estimate the joint PDF of $\mathbf{T}_1$ and $\mathbf{T}_2$ under $\mathcal{H}_0$ [8]. We assume that $p_{\mathbf{T}_1,\mathbf{T}_2}(\mathbf{t}_1, \mathbf{t}_2; \mathcal{H}_0)$ is completely known. However, under $\mathcal{H}_i$ $(i = 1, 2, \ldots, M)$ when a signal is present, we may not even have enough training samples to accurately estimate the marginal PDFs under $\mathcal{H}_i$. This is especially the case in the radar scenario, where the target is present for only a small portion of the time.

This work was supported by the Air Force Office of Scientific Research through the Air Force Research Laboratory Sensors Directorate under contract number FA8650-08-D-1303.

Hence, we want to construct appropriate joint PDFs under each $\mathcal{H}_i$ with as much information we have as possible, and make a classification using the constructed PDFs. A simple illustration is shown in Figure 1. Note that the result in this paper can be easily extended to the general multiple-sensor case.

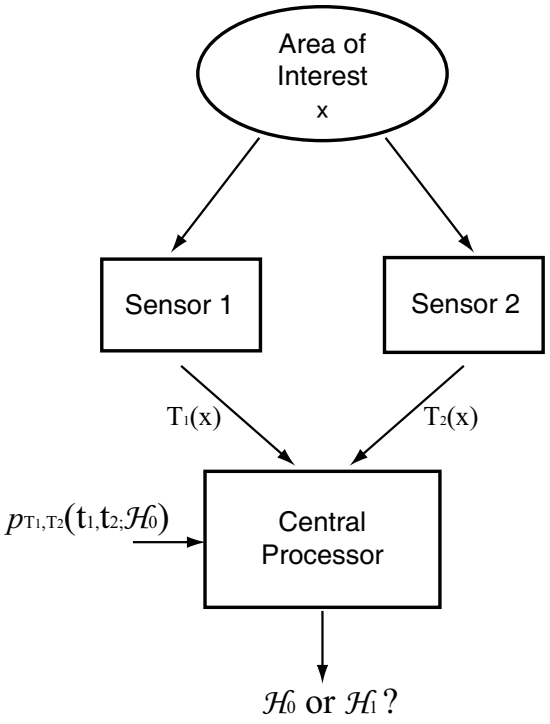

Fig. 1. Distributed classification system with two sensors.

III. Joint PDF Construction and Its Application in Classification

Since $p_{\mathbf{T}_1,\mathbf{T}_2}(\mathbf{t}_1,\mathbf{t}_2;\mathcal{H}_0)$ is the only information available, in order to specify the joint PDF $p_{\mathbf{T}_1,\mathbf{T}_2}(\mathbf{t}_1,\mathbf{t}_2;\mathcal{H}_i)$, we need the following assumptions [9].

1) The signal is small under each $\mathcal{H}_i$ and hence $p_{\mathbf{T}_1,\mathbf{T}_2}(\mathbf{t}_1,\mathbf{t}_2;\mathcal{H}_i)$ is close to $p_{\mathbf{T}_1,\mathbf{T}_2}(\mathbf{t}_1,\mathbf{t}_2;\mathcal{H}_0)$.

2) Under each $\mathcal{H}_i$, the joint PDF can be parameterized by some signal parameters $\boldsymbol{\theta}_i$ so that

$$p_{\mathbf{T}_1,\mathbf{T}_2}(\mathbf{t}_1,\mathbf{t}_2;\mathcal{H}_i) = p_{\mathbf{T}_1,\mathbf{T}_2}(\mathbf{t}_1,\mathbf{t}_2;\boldsymbol{\theta}_i)$$

$$p_{\mathbf{T}_1,\mathbf{T}_2}(\mathbf{t}_1,\mathbf{t}_2;\mathcal{H}_0) = p_{\mathbf{T}_1,\mathbf{T}_2}(\mathbf{t}_1,\mathbf{t}_2;\mathbf{0})$$

Hence the classification problem is to choose from

$$\mathcal{H}_i: \quad \boldsymbol{\theta} = \boldsymbol{\theta}_i \qquad \text{for } i = 1,\ldots,M$$

Let

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_1 \\ \mathbf{T}_2 \end{bmatrix}$$

so that the joint PDF $p_{\mathbf{T}_1,\mathbf{T}_2}(\mathbf{t}_1,\mathbf{t}_2;\boldsymbol{\theta}_i)$ can be written as $p_{\mathbf{T}}(\mathbf{t};\boldsymbol{\theta}_i)$. As shown in [9] with a first order Taylor expansion on the log-likelihood function under each $\mathcal{H}_i$, we can construct the PDF of $\mathbf{T}$ under $\mathcal{H}_i$ as

$$p_{\mathbf{T}}(\mathbf{t};\boldsymbol{\theta}_i) = \exp\left[\boldsymbol{\theta}_i^T\mathbf{t} - K(\boldsymbol{\theta}_i) + \ln p_{\mathbf{T}}(\mathbf{t};\mathbf{0})\right] \tag{1}$$

where

$$K(\boldsymbol{\theta}_i) = \ln E_0\left[\exp\left(\boldsymbol{\theta}_i^T\mathbf{T}\right)\right] \tag{2}$$

is the cumulant generating function of $p_{\mathbf{T}}(\mathbf{t};\mathbf{0})$, and it normalizes the PDF to integrate to 1. Note that it is assumed that $p_{\mathbf{T}}(\mathbf{t};\mathbf{0})$ is available or it can be estimated with reasonable accuracy.

In order to estimate the unknown parameters $\boldsymbol{\theta}_i$ in $p_{\mathbf{T}}(\mathbf{t};\boldsymbol{\theta}_i)$, we will use the maximum likelihood estimate (MLE) [10]. We see that in (1), the constructed PDF is in the form of an exponential family, and many nice properties are as follows:

1. $\mathbf{T}$ is a sufficient statistic for constructed PDF, and hence this PDF incorporates all the sensor information.

2. $K(\boldsymbol{\theta}_i)$ is convex by Holder's inequality [11]. Since maximizing $p_{\mathbf{T}}(\mathbf{t};\boldsymbol{\theta}_i)$ is equivalent to maximizing $\boldsymbol{\theta}_i^T\mathbf{t} - K(\boldsymbol{\theta}_i)$, this becomes a convex optimization problem and many existing methods can be readily utilized [12], [13].

3. It can be shown that by maximizing $p_{\mathbf{T}}(\mathbf{t};\boldsymbol{\theta}_i)$ over $\boldsymbol{\theta}_i$, the resulting PDF is asymptotically the closest to the true PDF $p_{\mathbf{T}}(\mathbf{t};\mathcal{H}_i)$ in KL divergence [9]. Similar arguments have been shown in [7], [14].

For classification, if we assume equal prior probabilities of each hypothesis, i.e., $p(\mathcal{H}_1) = p(\mathcal{H}_2) = \cdots = p(\mathcal{H}_M)$, the MAP rule can be reduced to the maximum likelihood (ML) rule [1]. When the MLE of $\boldsymbol{\theta}_i$ is found by maximizing $p_{\mathbf{T}}(\mathbf{t};\boldsymbol{\theta}_i)$ over $\boldsymbol{\theta}_i$, we consider $p_{\mathbf{T}}(\mathbf{t};\hat{\boldsymbol{\theta}}_i)$ as our estimate of $p_{\mathbf{T}}(\mathbf{t};\mathcal{H}_i)$ where $\hat{\boldsymbol{\theta}}_i$ is the MLE of $\boldsymbol{\theta}_i$. Hence similar to the ML rule, we will decide $\mathcal{H}_i$ for which the following is maximum over i:

$$p_{\mathbf{T}}(\mathbf{t};\hat{\boldsymbol{\theta}}_i) \tag{3}$$

By the monotonicity of the log function, we can equivalently decide $\mathcal{H}_i$ for which the following is maximum over i:

$$\ln\frac{p_{\mathbf{T}}(\mathbf{t};\hat{\boldsymbol{\theta}}_i)}{p_{\mathbf{T}}(\mathbf{t};\mathbf{0})} = \hat{\boldsymbol{\theta}}_i^T\mathbf{t} - K(\hat{\boldsymbol{\theta}}_i) \tag{4}$$

We will compare the performance of our classifier to that of the estimated MAP classifier. The estimated MAP classifier assumes that the PDF of $\mathbf{T}$ under $\mathcal{H}_i$ is known except for some unknown underlying parameters $\boldsymbol{\alpha}_i$. We still assume that $p(\mathcal{H}_1) = p(\mathcal{H}_2) = \cdots = p(\mathcal{H}_M)$. So the estimated MAP classifier finds the MLE of $\boldsymbol{\alpha}_i$ and chooses $\mathcal{H}_i$ for which the following is maximum over i:

$$p_{\mathbf{T}}(\mathbf{t};\hat{\boldsymbol{\alpha}}_i) \tag{5}$$

where $\hat{\boldsymbol{\alpha}}_i$ is the MLE of $\boldsymbol{\alpha}_i$. Note that for the estimated MAP classifier, $\boldsymbol{\alpha}_i$ are the unknown parameters in the true PDF under $\mathcal{H}_i$, while $\boldsymbol{\theta}_i$ are the unknown parameters in the constructed PDF under $\mathcal{H}_i$. Since the constructed PDF may or may not be the true PDF, the estimated MAP classifier assumes more information than our classifier.

IV. A Linear Model Example

Consider the following classification model:

$$\mathcal{H}_i: \quad \mathbf{x} = A_i \mathbf{s}_i + \mathbf{w} \tag{6}$$

where $\mathbf{s}_i$ is an $N \times 1$ known signal vector with the same length as $\mathbf{x}$, A_i is the unknown signal amplitude, and $\mathbf{w}$ is white Gaussian noise with known variance σ^2. Assume that instead of observing $\mathbf{x}$, we can only observe the measurements of two sensors

$$\begin{aligned} \mathbf{T}_1 &= \mathbf{H}_1^T \mathbf{x} \\ \mathbf{T}_2 &= \mathbf{H}_2^T \mathbf{x} \end{aligned} \tag{7}$$

where $\mathbf{H}_1$ is $N \times p_1$ and $\mathbf{H}_2$ is $N \times p_2$. Here p_1 and p_2 are the length for vectors $\mathbf{T}_1$ and $\mathbf{T}_2$ respectively. We can write (7) as

$$\mathbf{T} = \mathbf{G}^T \mathbf{x} \tag{8}$$

by letting

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_1 \\ \mathbf{T}_2 \end{bmatrix}$$

and

$$\mathbf{G} = [\mathbf{H}_1 \ \mathbf{H}_2]$$

where $\mathbf{G}$ is $N \times (p_1 + p_2)$ with $p_1 + p_2 \leq N$. We assume that $\mathbf{G}$ has full column rank so that there are no redundant measurements of the sensors. Note that $\mathbf{G}$ can be any matrix with full column rank.

Let $\mathcal{H}_0$ be the reference hypothesis when there is noise only, i.e.,

$$\mathcal{H}_0: \quad \mathbf{x} = \mathbf{w} \tag{9}$$

Since $\mathbf{x}$ is Gaussian under $\mathcal{H}_0$, according to (8), $\mathbf{T}$ is also Gaussian and

$$\mathbf{T} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{G}^T \mathbf{G}\right)$$

under $\mathcal{H}_0$. We construct the PDF under $\mathcal{H}_i$ as in (1) with

$$K(\boldsymbol{\theta}_i) = \ln E_0\left[\exp\left(\boldsymbol{\theta}_i^T \mathbf{T}\right)\right] = \frac{1}{2}\sigma^2 \boldsymbol{\theta}_i^T \mathbf{G}^T \mathbf{G} \boldsymbol{\theta}_i \tag{10}$$

The next step is to find the MLE of $\boldsymbol{\theta}_i$. Note that the MLE of $\boldsymbol{\theta}_i$ is found by maximizing ${\boldsymbol{\theta}_i}^T \mathbf{t} - K(\boldsymbol{\theta}_i)$ over $\boldsymbol{\theta}_i$. If this optimization procedure is carried without any constraint, then $\hat{\boldsymbol{\theta}}_i$ would be the same for all i. Hence we need some implicit constraints in finding the MLE. Since $\boldsymbol{\theta}_i$ represents the signal under $\mathcal{H}_i$, we should have

$$\boldsymbol{\theta}_i = A_i \mathbf{G}^T \mathbf{s}_i = E_{\mathcal{H}_i}(\mathbf{T}) \tag{11}$$

which is the mean of $\mathbf{T}$ under $\mathcal{H}_i$. As a result of (10), the MLE of $\boldsymbol{\theta}_i$ is found by maximizing

$$\boldsymbol{\theta}_i^T \mathbf{t} - K(\boldsymbol{\theta}_i) = \boldsymbol{\theta}_i^T \mathbf{t} - \frac{1}{2}\sigma^2 \boldsymbol{\theta}_i^T \mathbf{G}^T \mathbf{G} \boldsymbol{\theta}_i \tag{12}$$

with the constraint in (11). In this case, this is equivalent to finding the MLE of A_i. It can be found that

$$\hat{A}_i = \frac{\mathbf{s}_i^T \mathbf{G} \mathbf{t}}{\sigma^2 \mathbf{s}_i^T \mathbf{G} \mathbf{G}^T \mathbf{G} \mathbf{G}^T \mathbf{s}_i} \tag{13}$$

and

$$\hat{\boldsymbol{\theta}}_i = \frac{\mathbf{G}^T \mathbf{s}_i \mathbf{s}_i^T \mathbf{G} \mathbf{t}}{\sigma^2 \mathbf{s}_i^T \mathbf{G} \mathbf{G}^T \mathbf{G} \mathbf{G}^T \mathbf{s}_i} \tag{14}$$

Hence by removing the constant factors, the test statistic of our classifier for $\mathcal{H}_i$ is

$$\frac{(\mathbf{s}_i^T \mathbf{G} \mathbf{t})^2}{(\mathbf{G}^T \mathbf{s}_i)^T \mathbf{G}^T \mathbf{G} (\mathbf{G}^T \mathbf{s}_i)} \tag{15}$$

Next we consider the estimate MAP classifier. In this case, we assume that we know

$$\mathbf{T} \sim \mathcal{N}\left(A_i \mathbf{G}^T \mathbf{s}_i, \sigma^2 \mathbf{G}^T \mathbf{G}\right) \quad \text{under } \mathcal{H}_1$$

So A_i is really the unknown parameter in the true PDF under $\mathcal{H}_i$. It can be found that

$$\hat{A}_i = \frac{\mathbf{s}_i^T \mathbf{G} \left(\mathbf{G}^T \mathbf{G}\right)^{-1} \mathbf{t}}{\mathbf{s}_i^T \mathbf{G} \left(\mathbf{G}^T \mathbf{G}\right)^{-1} \mathbf{G}^T \mathbf{s}_i} \tag{16}$$

By removing the constant terms, the test statistic of the estimated MAP classifier for $\mathcal{H}_i$ is

$$\frac{(\mathbf{s}_i^T \mathbf{G} \left(\mathbf{G}^T \mathbf{G}\right)^{-1} \mathbf{t})^2}{\left(\mathbf{G}^T \mathbf{s}_i\right) \left(\mathbf{G}^T \mathbf{G}\right)^{-1} \left(\mathbf{G}^T \mathbf{s}_i\right)} \tag{17}$$

Note that (13) and (16) are different because (13) is the MLE of A_i under the constructed PDF and (16) is the MLE of A_i under the true PDF.

V. Simulation Results

For the model in (6)

$$\mathcal{H}_i: \quad \mathbf{x} = A_i \mathbf{s}_i + \mathbf{w}$$

let $A_1 = 0.5$, $A_2 = 1$, $A_3 = 1$ and

$$\begin{aligned} s_1(n) &= \cos(2\pi f_1 n) + 1 \\ s_2(n) &= \cos(2\pi f_2 n) + 0.5 \\ s_3(n) &= \cos(2\pi f_3 n) \end{aligned}$$

where $n = 0, 1, \ldots, N-1$ with $N = 20$, and $f_1 = 0.17$, $f_2 = 0.28$, $f_3 = 0.45$. Let $p(\mathcal{H}_1) = p(\mathcal{H}_2) = p(\mathcal{H}_3) = 1/3$. Assume that there are three sensors, each with an observation matrix as follows respectively:

$$\begin{aligned} \mathbf{H}_1 &= \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T \\ \mathbf{H}_2 &= \begin{bmatrix} 1 & \cos(2\pi f_1) & \cdots & \cos\left(2\pi f_1 (N-1)\right) \\ 1 & \cos(2\pi f_2) & \cdots & \cos\left(2\pi f_2 (N-1)\right) \end{bmatrix}^T \\ \mathbf{H}_3 &= \begin{bmatrix} 1 & \cos\left(2\pi (f_3 + 0.02)\right) & \cdots \\ & \cos\left(2\pi (f_3 + 0.02)(N-1)\right) \end{bmatrix}^T \end{aligned}$$

Note that in $\mathbf{H}_3$, we set the frequency to $f_3 + 0.02$. This is the case when the knowledge of the frequency is not accurate.

The test statistics are used as in (15) and (17) for the two methods respectively. The probabilities of correct classification are plotted versus $\ln(1/\sigma^2)$ in Figure 2. We see that their performances appear to be the same, and probabilities of correct classification goes to 1 as $\sigma^2 \to 0$.

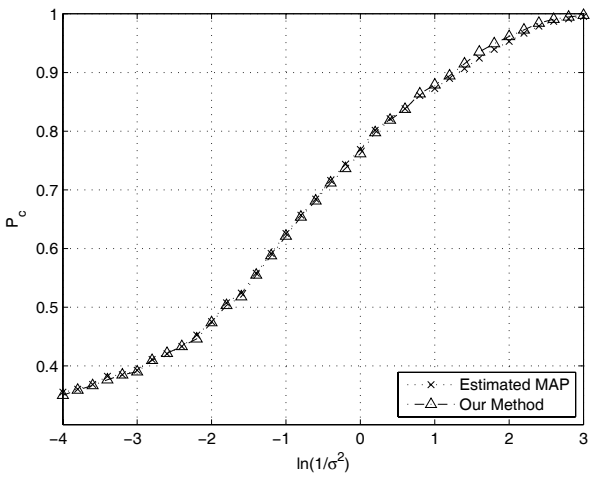

Fig. 2. Probability of correct classification for both methods.

VI. Conclusion

A novel method of constructing the joint PDF of sensor outputs for classification has been proposed. Only a reference PDF is needed in the construction. The constructed PDF is asymptotically the closest to the true PDF in KL divergence, and hence it asymptotically optimal. When applied to distributed classification, its performance is shown to be as good as the estimated MAP classifier, which assumes more information than our classifier.

References

[1] S. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[2] S. Thomopoulos, R. Viswanathan, and D. Bougoulias, "Optimal distributed decision fusion," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 25, pp. 761–765, Sep. 1989.

[3] Z. Chair and P. Varshney, "Optimal data fusion in multiple sensor detection systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 22, pp. 98–101, Jan. 1986.

[4] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, pp. 226–239, Mar. 1998.

[5] A. Sundaresan, P. Varshney, and N. Rao, "Distributed detection of a nuclear radioactive source using fusion of correlated decisions," in *Information Fusion, 2007 10th International Conference on*, 2007, pp. 1–7.

[6] S. Iyengar, P. Varshney, and T. Damarla, "A parametric copula based framework for multimodal signal processing," in *ICASSP*, 2009, pp. 1893–1896.

[7] S. Kay and Q. Ding, "Exponentially embedded families for multimodal sensor processing," in *ICASSP*, Mar. 2010.

[8] S. Kay, A. Nuttall, and P. Baggenstoss, "Multidimensional probability density function approximations for detection, classification, and model order selection," *IEEE Trans. Signal Process.*, vol. 49, pp. 2240–2252, Oct. 2001.

[9] S. Kay, Q. Ding, and D. Emge, "Joint pdf construction for sensor fusion and distributed detection," in *International Conference on Information Fusion*, Jun. 2010.

[10] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[11] L. Brown, *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, 1986.

[12] S. Boyd and L.Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[13] D. Luenberger, *Linear and Nonlinear Programming*, 2nd ed. Springer, 2003.

[14] S. Kay, "Exponentially embedded families - new approaches to model order estimation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 41, pp. 333–345, Jan. 2005.