Multilevel Data Integration with Applications in Sensor Networks

James C. Spall*
The Johns Hopkins University, Applied Physics Laboratory, Laurel, MD, 20723, USA

Long Wang†
Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, 21218, USA

This work considers a general system composed of multiple subsystems, where data can be collected at different levels of the system with possibly different probability distributions. Using the general formulations of the distributions and the principle of maximum likelihood estimation, we develop a method for estimating parameters, including uncertainty bounds on the estimates, associated with the relevant performance metric. The proposed method is demonstrated in two applications: (1) Detecting target locations by integrating data from both UAVs (unmanned aerial vehicles) and Doppler radar and (2) Detecting the position of "markers" in a problem of aerial refueling.

I. Introduction

The quantity and diversity of data on many types of defense systems are increasing with recent developments in data collection devices, such as mobile sensors and handheld computers. Consider a general stochastic system composed of multiple subsystems (or components), where the full system and each of the subsystems can generate various output data following some probability distributions. Such a system has a broad range of applications in practice, including reliability assessment in weapon systems, link loss and packet latency estimation in network tomography, attack detection and localization in cyber-physical systems, and travel time estimation in transportation networks (see [1] for a brief review). While the full system outputs can reflect the general information of the entire system, the outputs from each subsystem also reflect some partial information of the system. That is, while the full system outputs alone can provide some level of estimates for the parameter of interest at the full system level, being able to use all the data from both the full system and the subsystems can significantly improve the overall accuracy of the estimates. A schematic of the setting is given in Fig. 1. This paper demonstrates how the integration of data from different sources can be used in sensor networks. We illustrate the generic approach in two applications: (1) Detecting target locations by integrating data from both UAVs (unmanned aerial vehicles) and Doppler radar and (2) Detecting the position of "markers" in a problem of aerial refueling.

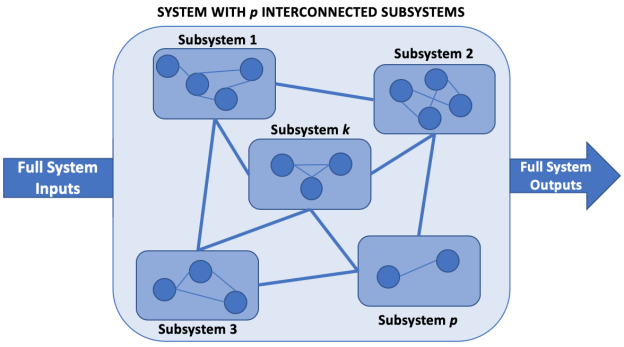

Fig. 1 Schematic of system with p subsystems.

*Principal Professional Staff, Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel, MD 20723, USA.
†Department of Applied Mathematics and Statistics, 3400 North Charles Street, Baltimore, MD, 21218, USA.

There are three key challenges in integrating data from different sources: i) the probability distributions for the different sources may differ from each other; ii) the sample sizes for each data source might be different and there may be no samples at all for some subsystems; and iii) the relationship between the full system and the subsystems might be complicated and it might be difficult or impossible to find an explicit mathematical function that relates the parameters of the probability distributions to each other. The method here addresses all of these challenges in a formal way to ensure that the analytical solution is portable across platforms while maintaining a specified level of performance.

II. Application

Let us describe in more detail the first of the applications (UAVs and Doppler) mentioned above. The second application (aerial tanker) has similar characteristics, as illustrated in Section III below.

In this work, we consider a specific sensor network application similar to [2], where multiple UAVs and a Doppler radar are deployed to simultaneously locate multiple targets. Because a Doppler radar can cover a wide range of area, it is viewed as the full system in our proposal framework. The UAVs are considered as the subsystems, because they can only detect a small neighborhood around themselves and are required to cooperate with each other in order to identify the location of the targets. Similar to [2], the UAVs here generate noisy position measurements about their nearest objects, including the targets or the other UAVs, and are allowed to communicate with each other. Specific calculations related to the radii of the orbits as well as to the covariance matrices of the measurement noises are provided in [2]. Nonetheless, there are some significant differences between [2] and our approach. For example, in [2], the multiple UAVs can only fly in some pre-defined orbits around a fixed central Doppler radar; there is no such restriction in the approach here. We also allow for greater flexibility in the type of data that are collected and we provide a formal means for uncertainty calculations.

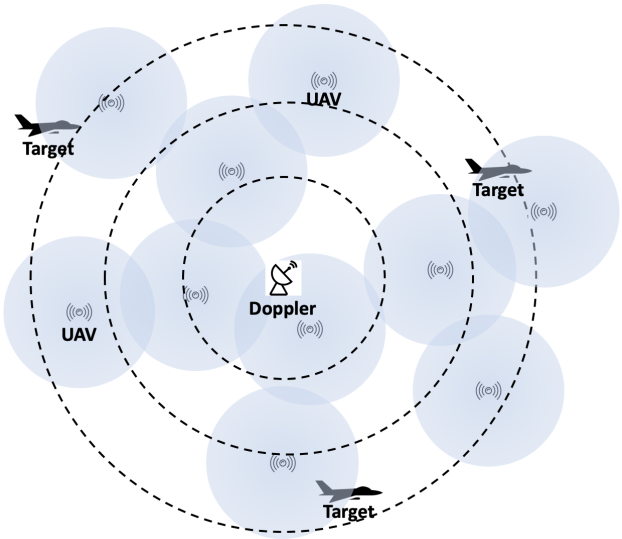

Fig. 2 Illustration of the spatial searching problem in a sensor network.

The general motivation here is to integrate the signals from both the long-range (i.e., Doppler radar) and short-range (i.e., UAVs) sensors to provide more accurate estimates of location. Figure 2 illustrates the general setting of this spatial searching problem. The Doppler radar is installed at the ground level and it can generate noisy location measurements of the targets. Although using only the Doppler measurements alone can provide some level of estimates of the target locations, it is often deemed as expensive and less accurate due to a limited number of measurements or a relatively high level of noise presented in the measurements. The individual UAVs, on the other hand, are more affordable and reliable since they are free to move in space and may potentially be very close to the targets, leading to a much smaller measurement uncertainty. The shortcoming of the UAV measurements is that the information only reflects the relative positions between the targets and the UAVs themselves. Therefore, we design a sensor network to allow the UAVs to share information in order to construct the estimations of the target locations. Depending on the location of the

targets, not all of the UAVs are active or reporting information about the targets. However, it is possible to optimize the deployment the UAVs such that the total coverage area is maximized. If more types of information about the targets are collected, this framework is also capable of estimating other properties of the targets, such as velocities and turning rates.

The idea of sensor cooperation, analogous to the UAVs here, is also discussed in [3–5], where several sensors are cooperatively establishing confident estimates of positions through assumptions, sequential verification, and iterative refinements to create an accurate map of the entire network. The approach in [3] and [4], however, is restricted to a static target. In [5], the UAV measurements can only reflect the relative positions among themselves, which can then be combined and used to estimate the exact positions of each individual UAV in the search area. In contrast, in our approach, we have additional information from the Doppler radar and our interest is to estimate the target positions. Because the Doppler radar measurements overlap with the UAV measurements, we must design a method to properly integrate all information into the overall sensor network in order to best estimate the target positions. That is, in our framework, the relative position measurements from the multiple UAVs are combined to not only reflect information about the sensors in the overall detection area but also the position information of the targets.

III. Overview of Analytical Approach and Extensions

To rigorously study the integration of multilevel data from the subsystems and the full system, we develop a method for estimating parameters—including uncertainty bounds on the estimates—associated with the relevant performance metric. We rely on analytical results in [1] and [6, chapter 1]. The work in [1] and [6] extends in a significant way the approach of [7], which was restricted to the case of binary-output subsystems and scalar parameters for the system and subsystems. The data on the full system and subsystems are allowed to follow different general exponential family distributions that depend on multiple parameters. The exponential family covers many of the most common distributions. Examples include chi-squared distributions (which are widely used in quality and reliability engineering), normal distributions (which are often assumed for modeling build-up of tolerances or life distribution of high-stress subsystems), and Weibull distributions (which are commonly used in applications such as failure time of subsystems subjected to fatigue, scheduling inspection or preventive maintenance activities). At the subsystem level, the distribution assumption allows not only the traditional binary ("0 or 1") outputs but also other non-binary type outputs that follow distributions such as normal or log-normal. In the binary outputs case, subsystem parameters often pertain to the success probabilities for the subsystems.

Using the general formulations for the likelihood function and the score vector (gradient vector of the log-likelihood function) that applies across all problem settings, this work can accommodate different arrangements of each subsystem within the full system by varying only the constraints between the full system and subsystem parameters. Based on the principles of maximum likelihood estimation (MLE), we derive a formal and easy to implement MLE formulation for the general structural relationship among the full system and each subsystem parameters. A key contribution of this work is the convergence theory related to consistency and asymptotic normality for the MLEs of the unknown full system and subsystem parameters. With either the asymptotic normality or bootstrap sampling, the MLE approach is well suited to capture the covariance between the parameter estimates and to provide asymptotic or finite-sample uncertainty bounds. The Fisher information matrix derived in the proof of asymptotic normality can be used to determine if the model parameters are identifiable (i.e., able to be estimated) and to determine the optimal number of full system and subsystem tests given cost and other constraints in the experimental budget.

Under the current framework, the results of this work may also be extended to the case where the full system and subsystem parameters may be linked to time-varying inputs, feedback controllers, or other effects such as system aging. Allowing the dynamics in this framework gives the proposed method the ability to solving many tracking problems with multiple detectors.

IV. Numerical Studies

A. Target Detection

The following numerical study is based on the framework in Section III above and the work in [8]. We consider the multiple UAVs as the subsystems, where each individual UAV can only detect a small neighborhood of itself and needs to work with other UAVs to locate the target. One potential motivation to have local UAVs is to make them more hidden than the broad signal, which can help to avoid revealing too much source information. Obviously, the fewer signals

the better in real operations due to security issues. Similar to the full system outputs, assume the subsystem outputs also follow the multivariate normal distributions, i.e., $\boldsymbol{X}^{(j)} = [X_1^{(j)}, X_2^{(j)}, X_3^{(j)}]^T \sim \mathcal{N}(\boldsymbol{\mu}^{(j)}, \boldsymbol{\Sigma}^{(j)})$, which measures the difference of the two nearest UAV positions or the difference of the target and its nearest UAV positions in the Cartesian coordinate system. Although UAVs are (of course) allowed to move in space, if the time scale of the sampling is much smaller than the time scale of the movement dynamics, we may consider the positions of the UAVs as fixed during the short sampling period. For example, denote the locations of two nearest UAVs as $[l_1^{(1)}, l_2^{(1)}, l_3^{(1)}]$ and $[l_1^{(2)}, l_2^{(2)}, l_3^{(2)}]$, respectively. Then, we have $\mu_1^{(1)} = l_1^{(2)} - l_1^{(1)}$, $\mu_2^{(1)} = l_2^{(2)} - l_2^{(1)}$ and $\mu_3^{(1)} = l_3^{(2)} - l_3^{(1)}$. Figure 3 illustrates the framework of this location detection problem with three UAVs and one Doppler radar. Note that we consider the measurements of the different locations between the UAVs as the subsystem measurements, not the locations of the UAVs themselves. This is why Fig. 3 contains three UAVs, but shows four different $\boldsymbol{X}^{(j)}$ for $j = 1, \ldots, 4$. Note that depending on the location of the target, not all the UAVs are active or reporting information about the target. For simplicity, assume that within the detection range of every UAV, there is only one other UAV in order to maximize the total coverage area.

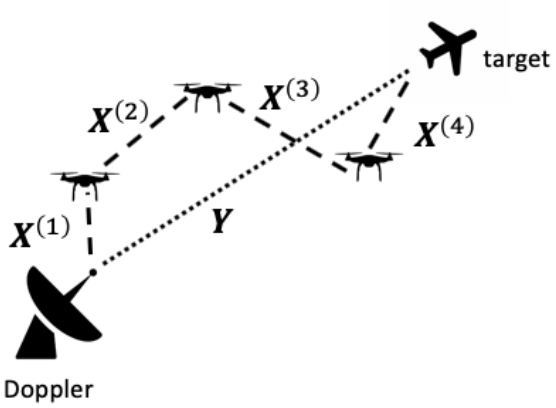

Fig. 3 Conceptual illustration of the sensor network: Doppler radar and UAVs are combined to provided information on target.

Under the Cartesian coordinate system, since the Doppler radar measurements can be constructed as $\boldsymbol{Y} = \sum_{j=1}^{4} \boldsymbol{X}^{(j)}$, we link the full system and subsystem parameter by defining the system structure function as $\boldsymbol{\mu}^F = \sum_{j=1}^{4} \boldsymbol{\mu}^{(j)}$ and $\boldsymbol{\Sigma}^F = \sum_{j=1}^{4} \boldsymbol{\Sigma}^{(j)}$. A similar construction can be found in [2] with the difference being in how UAVs are allowed to communicate with each other and whether the UAVs are fixed on predetermined orbits. Given a total of n_F measurements for the full system and n_j test measurements for the j-th subsystem that are independent from each other, the overall likelihood function has the form

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) = &-\frac{1}{2}\sum_{k=1}^{n_F}(\boldsymbol{Y}_k - \boldsymbol{\mu}^F)^T\boldsymbol{\Sigma}^F(\boldsymbol{Y}_k - \boldsymbol{\mu}^F) - \frac{n_F}{2}\log(\det\boldsymbol{\Sigma}^F)\\ &-\frac{1}{2}\sum_{j=1}^{4}\sum_{i=1}^{n_j}(\boldsymbol{X}_i^{(j)} - \boldsymbol{\mu}^{(j)})^T\boldsymbol{\Sigma}^{(j)}(\boldsymbol{X}_i^{(j)} - \boldsymbol{\mu}^{(j)})\\ &-\frac{1}{2}\sum_{j=1}^{4} n_j\log(\det\boldsymbol{\Sigma}^{(j)}) + \text{constant}.\end{aligned}$$

We consider a case with four subsystems and assume the following true parameter values:

$$\begin{aligned}\mu^{*(1)} &= [-1.2, 2.4, 6.2]^T, \mu^{*(2)} = [3.1, -0.4, 2.4]^T,\\ \mu^{*(3)} &= [0.1, 0.2, 2.1]^T, \mu^{*(4)} = [-0.2, -0.3, 3.2]^T,\end{aligned}$$

$$\boldsymbol{\Sigma}^{*(1)} = \boldsymbol{\Sigma}^{*(2)} = \boldsymbol{\Sigma}^{*(3)} = \boldsymbol{\Sigma}^{*(4)} = \begin{bmatrix} 10 & 0.8 & 0.8 \\ 0.8 & 10 & 0.8 \\ 0.8 & 0.8 & 10 \end{bmatrix}.$$

The true position values $\boldsymbol{\mu}^{*(j)}$ are chosen such that the distances between the Doppler radar and the active UAVs are about 6km, 8km, and 10km, respectively, matching the orbit radii assumptions in the numerical study of [2]. The distance between the Doppler radar and the target is then chosen to be 12km. The covariance matrices $\boldsymbol{\Sigma}^{*(j)}$ are also similar to the measurement noise assumptions for UAVs in [2].

Figure 4 shows the confidence interval for the target position under $n_F = 10$ for the full system measurements and various subsystem measurements, where for simplicity we have assumed that all the subsystems have the same number of measurements, i.e., $n_1 = n_2 = n_3 = n_4$. The solid lines represent the true parameter values for $\mu_1^{*F} = 1.8, \mu_2^{*F} = 1.9$ and $\mu_3^{*F} = 13.9$, where $[\mu_1^{*F}, \mu_2^{*F}, \mu_3^{*F}]^T = \boldsymbol{\mu}^{*F} = \sum_{j=1}^{4} \boldsymbol{\mu}^{*(j)}$. The dotted lines represent the empirically derived 95 percent confidence intervals for the true target position based on 500 independent replicates of the estimation process.

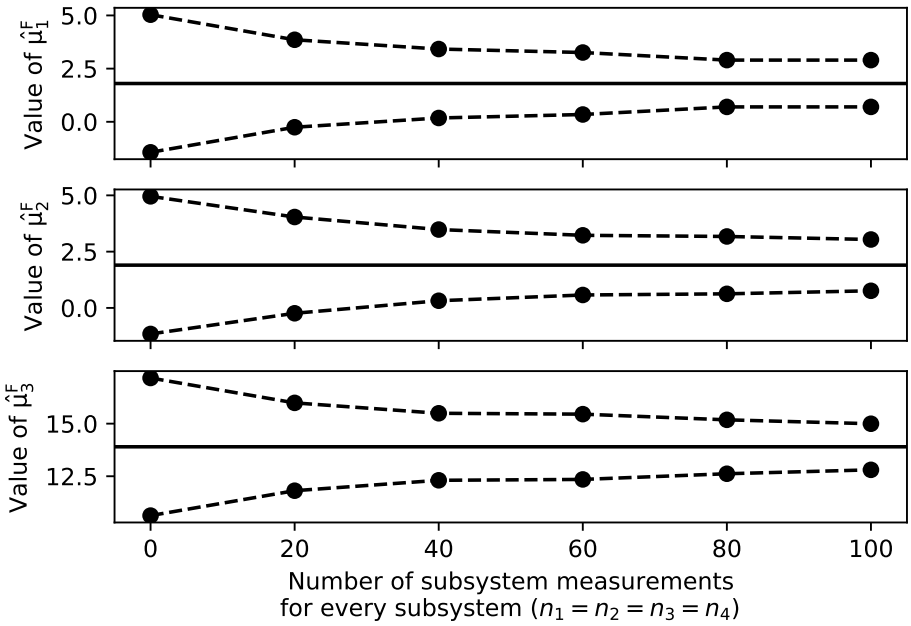

Fig. 4 95 percent confidence intervals for the true target position based on 500 independent replicates. All plots assume $n_F = 10$.

Based on Fig. 4, it is clear that by combining all the data from different sources, the estimate of the target position becomes more accurate as reflected in the fact that the confidence intervals are narrower. As more data are collected, the final estimates are converging towards the true parameter value and the likelihood of successfully identifying the target is significantly improved. The widths of the confidence intervals for the three mean values in Fig. 4 are reduced by 65, 63, and 66 percent, respectively, as the n_j increase from 0 to 100.

B. Markers Detection in Aerial Refueling

Another application is based on [9], where the goal is to detect the position of seven markers on the ground by integrating data from two UAVs: a tanker UAV and a receiver UAV. The markers are used to calibrate the positions of the UAVs for aerial refueling, where the point-matching operation is accomplished via the binocular LHM algorithm [10]. Here, we consider a simpler version of detecting only one marker on the ground since there is no fundamental difference when detecting a different number of markers in terms of the general methodology (see Fig. 5). The receiver UAV flying at a lower altitude can collect noisy measurements $\boldsymbol{X}^{(1)} \sim \mathcal{N}(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}^{(1)})$ of the marker. The tanker UAV

flying at a higher altitude can collect noisy measurements $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$ of the marker as well as relative position measurements $\boldsymbol{X}^{(2)} \sim \mathcal{N}(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}^{(2)})$ with respect to the receiver UAV.

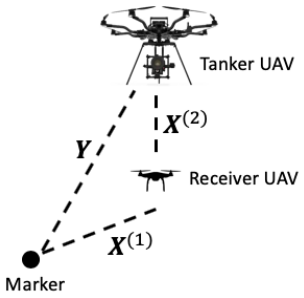

Fig. 5 Tanker UAV and receiver UAV are combined to provided information about the marker.

In our framework, $\boldsymbol{Y}$ corresponds to the full system measurements and $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ correspond to the subsystem measurements. Under the standard Cartesian coordinate system, we have $\boldsymbol{\mu}_Y = \boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)}$. Assume the following true parameter values

$$\boldsymbol{\mu}^{(1)} = [20, 13]^T, \boldsymbol{\mu}^{(2)} = [2, -4]^T,$$
$$\boldsymbol{\Sigma}^{(1)} = \boldsymbol{\Sigma}^{(2)} = \boldsymbol{\Sigma}_Y = \begin{bmatrix} 3 & 0.1 \\ 0.1 & 3 \end{bmatrix}.$$

We compute the distance in relevant units between the estimated and the true marker positions under various full system sample sizes n_F and subsystem sample sizes n_1 and n_2. (Distance is computed using the standard Euclidean norm.) The results are presented in Table 1 below and it is clear that as more data are collected the estimated position becomes more accurate.

Table 1 Distance between the estimated and true marker positions.

	$n_1 = n_2 = 0$	$n_1 = n_2 = 50$	$n_1 = n_2 = 100$
$n_F = 10$	0.788	0.031	0.018
$n_F = 50$	0.039	0.016	0.004

V. Conclusion

In this work, we consider the general setting of a system of subsystems and discuss a method to formally integrate data from the different levels of a complicated system. That is, we consider the formal integration of data from both the full system and the subsystems. Using the maximum likelihood formulation, two applications for sensor networks are considered. Through the numerical studies, it is shown that, in contrast to using only full system data or only subsystem data, the parameter estimates related to target location can be significantly improved when both full system and subsystem data are used in the estimation process.

Acknowledgments

This work was partially supported by the JHU/APL IRAD program.

References

[1] Wang, L., and Spall, J. C., "Beyond the Identification of Reliability for System with Binary Subsystems," *Proceedings of the American Control Conference*, Seattle, WA, 2017, pp. 158–163.

[2] Sun, T., Xin, M., and Jia, B., "Distributed Estimation in General Directed Sensor Networks Based on Batch Covariance Intersection," *Proceedings of the American Control Conference*, Boston, MA, 2016, pp. 5492–5497.

[3] Hernández, K., "Combined Sensor Information for Detection," *Proceedings of the Annual Conference on Information Sciences and Systems*, Baltimore, MD, 2015, pp. 1–6.

[4] Hernández, K., and Spall, J. C., "System Identification for Multi-sensor Data fusion," *Proceedings of the American Control Conference*, Chicago, IL, 2015, pp. 3931–3936.

[5] Savarese, C., Rabaey, J. M., and Beutel, J., "Location in Distributed Ad-hoc Wireless Sensor Networks," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4, Salt Lake City, UT, 2001, pp. 2037–2040.

[6] Wang, L., "Advances in System Identification and Stochastic Optimization," Ph.D. thesis, Johns Hopkins University, 2021.

[7] Spall, J. C., "Identification for Systems with Binary Subsystems," *IEEE Transactions on Automatic Control*, Vol. 59, No. 1, 2014, pp. 3–17.

[8] Wang, L., and Spall, J. C., "Multilevel Data Integration with Application in Sensor Networks," *Proceedings of the American Control Conference*, Denver, CO, 2020, pp. 5213–5218.

[9] Duan, H., Xin, L., and Chen, S., "Robust Cooperative Target Detection for a Vision-Based UAVs Autonomous Aerial Refueling Platform via the Contrast Sensitivity Mechanism of Eagle's Eye," *IEEE Aerospace and Electronic Systems Magazine*, Vol. 34, No. 3, 2019, pp. 18–30.

[10] Lu, C. P., Hager, G. D., and Mjolsness, E., "Fast and Globally Convergent Pose Estimation from Video Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 6, 2000, pp. 610–622.