Parameter Estimation from Heterogeneous/Multimodal Data Sets

Inbar Fijalkow, *Senior Member, IEEE*, Elad Heiman, and Hagit Messer, *Fellow, IEEE*

***Abstract*—Optimal parameter estimation requires simultaneous processing of all available measurements. The complexity of this task may become too large when measurements from two or more multimodal sensor networks are avaliable. In such cases, fusion of estimates obtained from each data set separately may be practical. In this paper, we derive the optimal linear combination of the possibly non-linear estimators, and propose sub-optimal weightings. We analyze the asymptotic performance gain of the first suboptimal approach with respect to the individual optimal estimates. The theoretical results are supported by simulations.**

***Index Terms*—Estimation theory, fisher information matrix, heterogeneous sensor networks, multimodal sensors.**

I. Introduction

With the recent development of sensor networks, signal processing applications may use data sets coming from very heterogeneous kind of sensors, [1]. For instance, to measure precipitation, Messer and Sendik propose to use both commercial telecommunication links (see [2]) and more classical rain gauges, [3]. In the first data set, the sensors observe a non-linear function of the rain parameters. These sensors are not very precise yet provide many samples in time and space, whereas the rain gauges are very precise but scarce in both time and space. To get the best tempo-spatial mapping of rainfall, important for various applications, it is of great interest to use all available data sets in an efficient way. While the optimal use of all available measurements for parameter estimation is to set the maximum likelihood (ML) estimate on their joint statistics, such a process maybe unrealistic either due to complexity or due to other practical constrains, such as the need for synchronization and large memory. In particular, in sensor networks, where different data sets may represents separate networks, the usage of the estimates provided by each network may be interesting. Fusion of estimators from multiple data sets is considered in [4]. However, the problem is addressed only when the parameters are observed through a linear data model. In [5] and [6], estimates' fusion is called data schrinkage and is also applied to covariance matrix estimation. In [7], schrinkage of one ML estimator to a target distribution is proposed to estimate the data entropy.

On the contrary of the literature, we address the estimation of parameters observed through at least one non-linear function. We use a linear combining of several estimators, as in the standart data fusion literature. However, if the linear combining of estmators is straightforward in the case of linear data models, it is not for non-linear data models. In particular, the optimal linear combining depends on the parameters to be estimated.

We discuss a sub-optimal and simple approach in which the best linear combination of the individual ML estimates is considered. The analytical study of the estimates' asymptotical variances results in an optimal combination defined by the individual Fisher information matrices. Upon normalization of the individual Fisher information matrices, we will discuss trends in the combination with respect to the different data set lengths and noise variances. Since the best combination value depends on the parameters to be estimated because of the non-linear function, we will look for a sub-optimal one and will provide an iterative simple algorithm to approach the optimal value.

The contributions of this letter are: (i) the proposition of two sub-optimal combination values, requiring no prior knowledge, given ML estimators of the parameters, (ii) the performance analysis of the optimal and first-suboptimal combiners with respect to the individual ML estimators.

The rest of this paper is organized as follows. Section II presents the data sets models, the individual ML estimates and introduces the linear combination. In Section III, we derive the optimal and sub-optimal linear combination. The respective performances of these estimators are studied and compared in terms of variances and covariance matrices in Section IV. Finally, we propose a numerical algorithm to approach the optimal value.

Notations: Bold letters stand for column vectors, capital letters for matrices. $E[]$ denotes the mean expectation. $Cov(\widehat{\boldsymbol{\theta}}) = E[(\widehat{\boldsymbol{\theta}} - E[\widehat{\boldsymbol{\theta}}])(\widehat{\boldsymbol{\theta}} - E[\widehat{\boldsymbol{\theta}}])^t]$ is the covariance matrix of the estimate $\widehat{\boldsymbol{\theta}}$. Its variance is defined as $var(\widehat{\boldsymbol{\theta}}) = tr(Cov(\widehat{\boldsymbol{\theta}}))$ where $tr()$ is the trace operator.

II. Multimodal Data Sets Model and Problem Setting

We consider D data sets induced by the $P-$dimensional parameter vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_P)^t$ to be estimated.

A. Individual Estimate $\widehat{\boldsymbol{\theta}}_i$

The N_i samples, $N_i \geq P$, of data set i are modeled as,

$$\mathbf{x}_i = \mathbf{s}_i(\boldsymbol{\theta}) + \mathbf{n}_i, i = 1, \dots, D. \tag{1}$$

Manuscript received November 23, 2015; revised January 21, 2016; accepted January 26, 2016. Date of publication January 29, 2016; date of current version February 15, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Steeve Zozor.

I. Fijalkow is with the ETIS/ENSEA, Université de Cergy-Pontoise, CNRS, F-95000 Cergy-Pontoise, France (e-mail: inbar.fijalkow@ensea.fr).

E. Heiman and H. Messer are with the School of Electrical Engineering, Tel Aviv University, Ramat Aviv 69978, Israel (e-mail: messer@eng.tau.ac.il).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2016.2523886

1070-9908 © 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Both data and parameters are taken as real valued. No prior knowledge is assumed on $\boldsymbol{\theta}$, so that $\mathbf{s}_i(\boldsymbol{\theta})$ is deterministic. Moreover, $\mathbf{s}_i()$ is assumed known. The observation noise $\mathbf{n}_i$ is assumed to follow a Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma_i^2 I_{N_i})$. $\mathbf{n}_i$ and $\mathbf{n}_j$ are assumed independent for $i \neq j$.

$\widehat{\boldsymbol{\theta}}_i$ denotes the ML estimate obtained from $\mathbf{x}_i$. The corresponding $P \times P$ error covariance matrix is denoted as $Cov(\widehat{\boldsymbol{\theta}}_i)$. Asymptotically with respect to the data set size and under the condition that $\mathbf{s}_i()$ is derivable twice, $Cov(\widehat{\boldsymbol{\theta}}_i)$ reaches the minimum value provided by the $P \times P$ Fisher information matrix (FIM) $F_i^{-1}(\boldsymbol{\theta})$, [8, p.167]. From (1),

$$F_i(\boldsymbol{\theta}) = \frac{1}{\sigma_i^2} G_i(\boldsymbol{\theta}) G_i(\boldsymbol{\theta})^t \tag{2}$$

with the $P \times N_i$ matrix $G_i(\boldsymbol{\theta})$ defined by its j^{th} row $\partial \mathbf{s}_i(\boldsymbol{\theta})^t / \partial \theta_j$, for $j = 1, \ldots, P$. All $G_i(\boldsymbol{\theta})$, $i = 1, \ldots, D$, are assumed to be full rank, equal to the number of parameters P.

B. *Estimation from Heterogeneous Data Sets*

We denote by $\widehat{\boldsymbol{\theta}}_{ML}$ the ML estimate obtained from the joint distribution of $\mathbf{x}_i$, $i = 1, \ldots, D$. From the independence between different $\mathbf{n}_i$ and their normality, the joint FIM and asymptotic ML covariance matrix is $(\sum_{i=1}^{D} F_i(\boldsymbol{\theta}))^{-1}$, see [8]. In extreme cases where one data set is much better than the others, either because its signal to noise ratio or its size the larger, the fused estimate should be mostly due to one estimate. In such cases, considering the optimal, ML estimate induces unneeded complexity. To reduce complexity, should we then consider only one data set and suppress the others? Alternatively, is it possible to improve the best individual estimate by a weighted averaging with the others?

In order to achieve a better complexity and performance trade-off than choosing the "best" data set, we suggest to look at $\alpha_i \in [0, 1]$ defining the following sub-optimal estimate,

$$\widehat{\boldsymbol{\theta}}_\alpha = \sum_{i=1}^{D} \alpha_i \widehat{\boldsymbol{\theta}}_i \text{ with } \sum_{i=1}^{D} \alpha_i = 1 \tag{3}$$

where $\widehat{\boldsymbol{\theta}}_i$, $i = 1, \ldots, D$ are the ML estimates from the i^{th} data set. The linear constraint $\sum_{i=1}^{D} \alpha_i = 1$ maintains (3) asymptotically unbiased. Since the variance of $\widehat{\boldsymbol{\theta}}_\alpha$ is a quadratic function of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)^t$, minimizing its variance under the linear constraint and convex set defined by $\alpha_i \in [0, 1]$, results from convex optimization [9] in a unique solution, denoted $\boldsymbol{\alpha}_*$. Unfortunatly, except for trivial cases, $\boldsymbol{\alpha}_*$ depends on the unknown parameters $\boldsymbol{\theta}$. The aim of this paper is to propose sub-optimal estimates that do not require the knowledge of $\boldsymbol{\theta}$, and that still improve the performances with respect to the individual estimates, $\widehat{\boldsymbol{\theta}}_i$.

TABLE I
CASES OF EXTREMELY HETEROGENEOUS DATASETS

$\sigma_1^2 \approx \sigma_2^2$, $N_2 >> N_1$ or $N_2 \approx N_1$, $\sigma_1^2 >> \sigma_2^2$	$\Rightarrow$	$\rho_1 >> 1$	$\Rightarrow$	$\widehat{\boldsymbol{\theta}}_{\alpha^*} \approx \widehat{\boldsymbol{\theta}}_2$
$\sigma_1^2 \approx \sigma_2^2$, $N_2 << N_1$ or $N_2 \approx N_1$, $\sigma_1^2 << \sigma_2^2$	$\Rightarrow$	$\rho_1 << 1$	$\Rightarrow$	$\widehat{\boldsymbol{\theta}}_{\alpha^*} \approx \widehat{\boldsymbol{\theta}}_1$

III. SUB-OPTIMAL ESTIMATES

A. *Optimal Value of* α

The $\mathbf{n}_i$ being decorrelated from each other, (3) yields to

$$\begin{aligned} var(\widehat{\boldsymbol{\theta}}_\alpha) &= \sum_{i=1}^{D} \alpha_i^2 var\left(\widehat{\boldsymbol{\theta}}_i\right) \\ &= \sum_{i=1}^{D-1} \alpha_i^2 tr\left(F_i^{-1}(\boldsymbol{\theta})\right) + \left(1 - \sum_{i=1}^{D-1} \alpha_i\right)^2 tr\left(F_D^{-1}(\boldsymbol{\theta})\right) \end{aligned} \tag{4}$$

where we replaced α_D by $1 - \sum_{i=1}^{D-1} \alpha_i$ to get ride of the equality constraint. We minimize the quadratic function (4) over the convex set defined by $0 \leq \alpha_i \leq 1$ for $i = 1, \ldots, D-1$ and $0 \leq 1 - \sum_{i=1}^{D-1} \alpha_i \leq 1$. Since our data sets have different sizes, we denote $\tilde{F}_i(\boldsymbol{\theta}) = \sigma_i^2 / N_i F_i(\boldsymbol{\theta})$ the normalized FIM with respect to the noise and to the data set size.

Proposition 1: The optimal weights for (3) are given by

$$\alpha_i^*(\boldsymbol{\theta}) = \frac{1}{\sum_{j=1}^{D-1} \frac{\rho_i \lambda_i^*(\boldsymbol{\theta})}{\rho_j \lambda_j^*(\boldsymbol{\theta})} + \rho_i \lambda_i^*(\boldsymbol{\theta})}, i = 1, \ldots, D-1 \tag{5}$$

with $\lambda_i^*(\boldsymbol{\theta}) = tr(\tilde{F}_i^{-1}(\boldsymbol{\theta})) / tr(\tilde{F}_D^{-1}(\boldsymbol{\theta}))$ and $\rho_i = \sigma_i^2 / \sigma_D^2 N_D / N_i$.

Proof: The Lagragian to be minimized is $var(\widehat{\boldsymbol{\theta}}_\alpha) - 2\mu(1 - \sum_{i=1}^{D-1} \alpha_i)$ with Lagrange multiplier $\mu \geq 0$. Solving the KKT equations [9]: $\alpha_i tr(\tilde{F}_i^{-1}(\boldsymbol{\theta})) - (1 - \sum_{i=1}^{D-1} \alpha_i)$ $tr(\tilde{F}_D^{-1}(\boldsymbol{\theta})) + \mu = 0$ for $i = 1, \ldots, D-1$ and $\mu(1 - \sum_{i=1}^{D-1} \alpha_i) = 0$, we get $\alpha_i = 1 / tr(\tilde{F}_i^{-1}(\boldsymbol{\theta})) / tr(\tilde{F}_D^{-1}(\boldsymbol{\theta}))$ $+ \sum_{j=1}^{D-1} tr(\tilde{F}_i^{-1}(\boldsymbol{\theta})) / tr(\tilde{F}_j^{-1}(\boldsymbol{\theta}))$ in $[0, 1]$. Introducing the definitions ρ_i and $\lambda_i^*(\boldsymbol{\theta})$ results in (5). ■

Except for the case of linear signal models, (5) requires the knowledge of $\boldsymbol{\theta}$ to be computed. We are looking for ways to choose these weights with no knowledge of $\boldsymbol{\theta}$. The case $D = 2$ gave us the idea of the heuristic proposed next.

Corollary 1: When $D = 2$, $\alpha_1^*(\boldsymbol{\theta}) = 1/1 + \rho_1 \lambda_1^*(\boldsymbol{\theta})$ is bounded in $[1/1 + \rho_1 \lambda_{max}(\boldsymbol{\theta}), 1/1 + \rho_1 \lambda_{min}(\boldsymbol{\theta})]$ where $\lambda_{max}(\boldsymbol{\theta})$ (resp. $\lambda_{min}(\boldsymbol{\theta})$) is the maximal (resp. minimal) eigenvalue of $\mathcal{F}(\boldsymbol{\theta}) = \tilde{F}_2^{1/2}(\boldsymbol{\theta}) \tilde{F}_1^{-1}(\boldsymbol{\theta}) \tilde{F}_2^{1/2}(\boldsymbol{\theta})$.

This comes from $\lambda_1^*(\boldsymbol{\theta}) = tr(\tilde{F}_1^{-1}(\boldsymbol{\theta})) / tr(\tilde{F}_2^{-1}(\boldsymbol{\theta})) = tr(\mathcal{F}(\boldsymbol{\theta}) \tilde{F}_2^{-1}(\boldsymbol{\theta})) / tr(\tilde{F}_2^{-1}(\boldsymbol{\theta}))$ and $tr(A) \lambda_{min}(B) \leq tr(AB) \leq tr(A) \lambda_{max}(B)$ for any semi-positive definite matrices A and B.

For some extreme cases of the data sets heterogeneousness, the optimal weighting in equation (3) is provided in Table I (by definition $\lambda_1^*(\boldsymbol{\theta}) \neq 0$).

B. *Sub-optimal Value of* α

Obviously, when the $\lambda_i^*(\boldsymbol{\theta})$ are of the order of 1, the values of ρ_i drive the weighting of $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\alpha}_*}$ in equation (5). Therefore, we suggest to set a value for $\boldsymbol{\alpha}$ independent of $\boldsymbol{\theta}$, as a function of the ρ_j, $j = 1, \ldots, D-1$ only. We propose

$$\alpha_i^h = \frac{1}{\sum_{j=1}^{D-1} \frac{\rho_i}{\rho_j} + \rho_i} \text{ for } i = 1, \ldots, D-1 \tag{6}$$

and $\alpha_D^h = 1 - \sum_{i=1}^{D-1} \alpha_i^h = 1/1 + \sum_{i=1}^{D-1} \rho_i$.

Although it is heuristic, $\boldsymbol{\alpha}^h$ preserves the desired properties of $\boldsymbol{\alpha}^*$ as in Table I. Moreover, when the size of each data set is very large, applying central limit theorem to the entries of $G_i(\boldsymbol{\theta})$, taken as random i.i.d., yields $F_i(\boldsymbol{\theta})$ in (2) to behave as $\sigma_i^2/N_i I_P$, [10]. Thus, $\boldsymbol{\alpha}^*(\boldsymbol{\theta}) \approx \boldsymbol{\alpha}^h$ for any $\boldsymbol{\theta}$ and large enough data sets. These assumptions are taken in this paragraph only as a motivation for (6).

IV. Performance Analysis

We study next the conditions under which $\boldsymbol{\alpha}^h$ improves the performance with respect to the individual ML estimates. The analysis is done in the asymptotic regime where the FIM describes the ML behavior. For the sake of simplicity, we consider in this section the case $D = 2$, and simplify the notation by omitting the index 1 in α, ρ and λ.

A. Variance Analysis

Thanks to the FIM bound and by definition of $\alpha^*(\boldsymbol{\theta})$ minimizing the variance of (3),

$$tr\left((F_1(\boldsymbol{\theta}) + F_2(\boldsymbol{\theta}))^{-1}\right) \leq var(\widehat{\boldsymbol{\theta}}_{\alpha^*(\boldsymbol{\theta})}) \leq var(\widehat{\boldsymbol{\theta}}_i), i = 1, 2,$$
$$\text{and } var(\widehat{\boldsymbol{\theta}}_{\alpha^*(\boldsymbol{\theta})}) \leq var(\widehat{\boldsymbol{\theta}}_{\alpha^h}).$$

Proposition 2 studies the conditions for $var(\widehat{\boldsymbol{\theta}}_{\alpha^h})$ to overcome the individual performances.

Proposition 2: Asymptotically, $var(\widehat{\boldsymbol{\theta}}_{\alpha^h}) \leq var(\widehat{\boldsymbol{\theta}}_i)$, $i = 1, 2$, if and only if $1/2 + \rho \leq \lambda^*(\boldsymbol{\theta}) \leq 2 + 1/\rho$.

Proof: For the sake of space, we omit the dependence in $\boldsymbol{\theta}$.

$$\begin{aligned} var\left(\widehat{\boldsymbol{\theta}}_\alpha\right) &= tr(F_2^{-1})\left(\alpha^2 tr\left(F_1^{-1}\right)/tr\left(F_2^{-1}\right) + (1-\alpha)^2\right) \\ &= tr(F_2^{-1})(\alpha^2\rho\lambda^* + (1-\alpha)^2) \end{aligned}$$

Simple calculations result in $var(\widehat{\boldsymbol{\theta}}_2) - var(\widehat{\boldsymbol{\theta}}_\alpha) = tr(F_2^{-1})\alpha(2 - \alpha(1 + \rho\lambda^*))$. It is positive if and only if $\lambda^* \leq 2/(\alpha\rho) - 1/\rho$. For $\alpha = \alpha^h$, the condition becomes $\lambda^* \leq 2 + 1/\rho$. In the same way, $var(\widehat{\boldsymbol{\theta}}_1) - var(\widehat{\boldsymbol{\theta}}_\alpha) = tr(F_2^{-1})(1-\alpha)((1+\alpha)\rho\lambda^* - (1-\alpha))$ is positive if and only if $\lambda^* \geq (1-\alpha)/(\rho(1+\alpha))$ which becomes $\lambda^* \geq 1/(2+\rho)$ for $\alpha = \alpha^h$. ■

Since $1/2 > 1/2 + \rho$ and $2 < 2 + 1/\rho$, $\lambda^*(\boldsymbol{\theta}) \in [1/2; 2]$ provide tighter bounds than that of Proposition 2, valid for any ρ. We deduce that when $1/2 \leq \lambda^*(\boldsymbol{\theta}) \leq 2$, the sub-optimal α^h in (6) allows $\widehat{\boldsymbol{\theta}}_{\alpha^h}$ to overcome the individual ML estimates.

B. Covariance Analysis

In the case of strictly more than 1 parameter, $P > 1$, the analysis of the variance may not be satisfactory enough so that we consider next the analysis of the covariance matrices. The FIM satisfies $(F_1(\boldsymbol{\theta}) + F_2(\boldsymbol{\theta}))^{-1} \leq Cov(\widehat{\boldsymbol{\theta}}_\alpha)$ for any α[1]. Yet, the ordering of the different covariance matrices involved is not always possible.

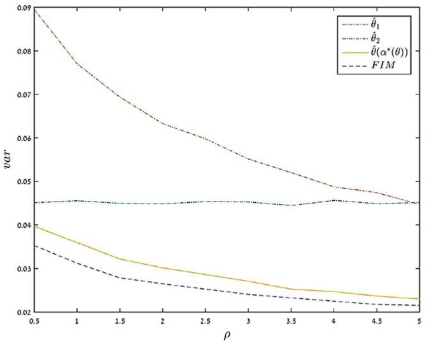

Fig. 1. Variance of $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \hat{\boldsymbol{\theta}}(\alpha^*(\boldsymbol{\theta}))$ and tr(FIM) with respect to ρ, $\boldsymbol{\theta} = (\pi/4, \pi/8)^t$.

Proposition 3: Asymptotically, a reduction of the covariance matrices of the linear-optimal and linear-suboptimal estimates with respect to the individual ML ones, is achieved under the following conditions:

- $Cov(\widehat{\boldsymbol{\theta}}_{\alpha^*(\boldsymbol{\theta})}) \leq F_i^{-1}(\boldsymbol{\theta})$ if and only if, for $k = 1, \ldots, P$, $(\lambda_k/\rho - 1)/(2\rho) \leq \lambda^*(\boldsymbol{\theta}) \leq 2/(\rho(1/\lambda_k - \rho))$.
- $Cov(\widehat{\boldsymbol{\theta}}_{\alpha^h}) \leq F_i^{-1}(\boldsymbol{\theta})$ if and only if, for $k = 1, \ldots, P$, $\rho/(1 + 2/\rho) \leq \lambda_k \leq (1 + 2\rho)\rho$.

Proof: Given the eigendecomposition of $\mathcal{F} = UDiag(\lambda_k)U^\dagger$,

$$\begin{aligned} Cov\left(\widehat{\boldsymbol{\theta}}_\alpha\right) &= \alpha^2 F_1^{-1} + (1-\alpha)^2 F_2^{-1} \\ &= F_2^{-1/2}\left(\alpha^2\mathcal{F} + (1-\alpha)^2 I\right) F_2^{-1/2} \\ &= F_2^{-1/2} UDiag\left(\alpha^2\lambda_k/\rho + (1-\alpha)^2\right) U^\dagger F_2^{-1/2} \end{aligned}$$

Therefore, $F_1^{-1} - Cov(\widehat{\boldsymbol{\theta}}_\alpha) = (1-\alpha)F_2^{-1/2}UDiag((1+\alpha)\lambda_k/\rho - (1-\alpha))U^\dagger F_2^{-1/2}$ is positive iff $(1+\alpha)\lambda_k/\rho - 1 + \alpha \geq 0$. $F_2^{-1} - Cov(\widehat{\boldsymbol{\theta}}_\alpha) = \alpha F_2^{-1/2}UDiag(2 - \alpha - \alpha\lambda_k/\rho)U^\dagger F_2^{-1/2}$ is positive iff $2 - \alpha - \alpha\lambda_k/\rho \geq 0$. Replacing α by $\alpha^*(\boldsymbol{\theta})$ (resp. α^h), we get the first (resp. second) line of proposal 3. ■

C. Numerical Illustration

We consider a simulation example with $P = 2$. Data set 1 is linear with respect to the parameters so that G_1 does not depend on $\boldsymbol{\theta}$, $tr(F_1^{-1}) = \sigma_1^2 tr((G_1 G_1^t)^{-1})$. We take $N_1 = 2$, $\sigma_1 = 0.05$ and $tr(F_1^{-1}) = 2$. For the second data set, we consider the non-linear, $\mathrm{s}_2(\boldsymbol{\theta}) = H_2(\sin(\theta_1) + \sin(\theta_2), \cos(\theta_1) + \cos(\theta_2))^t$ where H_2 spans the observations over N_2/P rows yielding to $tr(F_2^{-1}) = 2N_2\sigma_2^2/P\sin{(\theta_1 - \theta_2)}^2$ when $\sin(\theta_1 - \theta_2) \neq 0$. Note that $\alpha^*(\boldsymbol{\theta}) = 1$ otherwise. We set $\sigma_2 = 0.05$ and modify N_2 to vary ρ.

Fig. 1 shows the variances of the individual and linearly combined estimates with respect to ρ for $\boldsymbol{\theta} = (\pi/4, \pi/8)^t$. This value of $\boldsymbol{\theta}$ was chosen to satisfy all requirements in Propositions 2 and 3. As expected, we can verifiy on Fig. 1 that the best combined estimator $\widehat{\boldsymbol{\theta}}_{\alpha^*}$ outperforms the indivual

[1] For matrices, $A \leq B$ means $B - A$ is semi-positive definite.

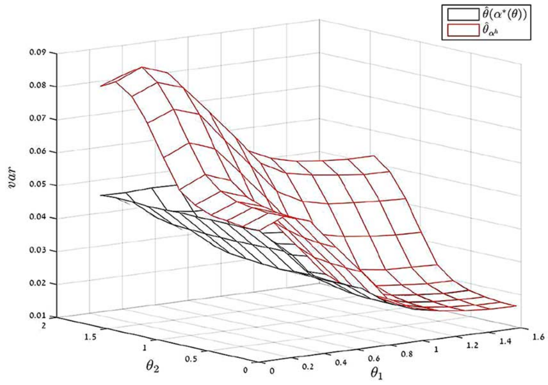

Fig. 2. Variances of $\widehat{\boldsymbol{\theta}}_{\alpha^*(\boldsymbol{\theta})}$ and $\hat{\boldsymbol{\theta}}_{\alpha^h}$ with respect to θ_1 and θ_2, $\rho = 1$.

Algorithm 1. Proposed iterative algorithm

```
1) Initialization: $\lambda_i^{(0)} = 1$ (implying $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\alpha}^h$)
2) Iteration $k$:
    • $\alpha_i^{(k)} = 1/\sum_{j=1}^{D-1} \rho_i \lambda_i^{(k)}/\rho_j \lambda_j^{(k)} + \rho_i \lambda_i^{(k)}, i = 1, \ldots, D-1$
    • $\widehat{\boldsymbol{\theta}}^{(k)} = \sum_{i=1}^{D-1} \alpha_i^{(k)} \widehat{\boldsymbol{\theta}}_i + (1 - \sum_{i=1}^{D-1} \alpha_i^{(k)}) \widehat{\boldsymbol{\theta}}_D$
    • $\lambda_i^{(k+1)} = tr(\tilde{F}_i^{-1}(\widehat{\boldsymbol{\theta}}^{(k)}))/tr(\tilde{F}_D^{-1}(\widehat{\boldsymbol{\theta}}^{(k)}))$
3) Repeat 2) until $|\lambda_i^{(k+1)} - \lambda_i^{(k)}| \leq \epsilon$.
```

estimates. The gain with respect to $\hat{\boldsymbol{\theta}}_1$ is especially high when ρ is small. For any ρ, the loss in performance of $\hat{\boldsymbol{\theta}}(\alpha^*(\boldsymbol{\theta}))$ with respect to the global ML remains small, showing the interest of optimizing α.

In the second simulation, we consider the same data sets as previously, with $N_2 = 2$ so that $\rho = 1$. Fig. 2 shows the variances of the linearly combined estimates for $\alpha = \alpha_\rho$ and $\alpha^*(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. For most values of $\boldsymbol{\theta}$, the two estimators behave quite similarly. However, when $\boldsymbol{\theta} \approx (0, \pi/2)^t$, the gap between $\widehat{\boldsymbol{\theta}}_{\alpha^*(\boldsymbol{\theta})}$ and $\widehat{\boldsymbol{\theta}}_{\alpha^h}$ becomes large (for this value of $\boldsymbol{\theta}$, $\alpha^*(\boldsymbol{\theta}) \approx 1$ whereas $\alpha^h \approx 1/3$). Thus, Fig. 2 validates the interest of approching $\alpha^*(\boldsymbol{\theta})$ rather than using α^h.

V. Numerical Computation of $\boldsymbol{\alpha}^*$

As we have seen in the previous example, one can be interested in approaching the unknown $\boldsymbol{\alpha}^*(\boldsymbol{\theta})$ numerically. To this end, we propose the following heuristic iterative algorithm. Algorithm 1 requires only the knowledge of ρ_i and the expression of $tr(\tilde{F}_i^{-1}(\boldsymbol{\theta}))$ for a given $\boldsymbol{\theta}$. If ρ_i are unknown, they could also be searched iteratively. To do so one should replace λ_i by $\rho_i \lambda_i$ and $\tilde{F}_i$ by F_i in the proposed algorithm. However, the convergence could be much slower, especially if $\rho_i \gg 1$ or $\rho_i \ll 1$. The convergence of the proposed fixed point algorithm depends on the numerical properties of the non-linear functions $tr(\tilde{F}_i^{-1}())$ and must be studied case by case, [11].

We consider the same simulation setting as in Section IV-C for $\boldsymbol{\theta} = (0, \pi/8)^t$ and $\rho = 2$ where $\alpha^h = 1/3$ and $\alpha^* = 0.7735$. The value of $\alpha^{(k)}$ is averaged over 500 random drawing of the data. Fig. 3 shows that the proposed numerical computation $\alpha^{(k)}$ improves with respect to the initial α^h since it is much closer to α^* within 4 iterations so that the induced numerical complexity is not very high.

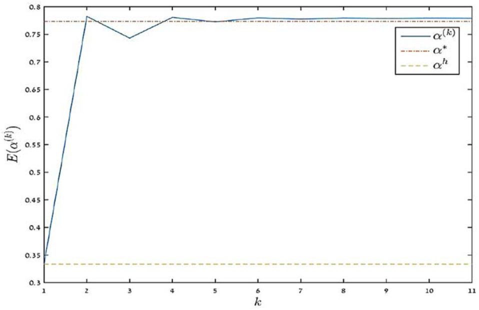

Fig. 3. Mean of $\alpha^{(k)}$ versus k. $\rho = 2, \boldsymbol{\theta} = (0, \pi/8)^t$.

VI. Conclusion

In the broad, complex field of multimodal data fusion from heterogeneous detectors [1], we focus our attention on an important special case. We study the case where a parameter vector $\boldsymbol{\theta}$ is observed by different independent sensors networks, each consisting of a known, possibly non-linear transformation on $\boldsymbol{\theta}$ in additive white Gaussian noise (1). In this case, we study a fused estimate of $\boldsymbol{\theta}$ where a linear combining of individual ML estimates is imposed. We derive the optimal combining coefficients which minimize the asymptotical variance of the combined estimate. While the optimal coefficients depend on the unknown parameter vector, we propose a sub-optimal combining which depends only on the noise levels and sizes of the data sets. We analyze the conditions under which it is guaranteed asymptotically that the performance of resulting estimate is better than each of the individual estimates, and we validate it by simulations. Our results provide an important tool for parameter estimation from big data, collected by independent sensor networks, since it is both simple to implement, and it guarantees improving estimation performance when adding data sets. We also could extend the problem by considering non independent addtive noise.

References

[1] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges and prospectives," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[2] O. Goldshtein, H. Messer, and A. Zinevich, "Rain rate estimation using measurements from commercial telecommunications links," *IEEE Trans. Signal Process.*, vol. 57, no. 4, Apr. 2009.

[3] H. Messer and O. Sendik, "A new approach to precipitation monitoring," *IEEE Signal Process. Mag.*, May 2015.

[4] X. R. Li, Y. Zhu, J. Wang, and C. Han, "Optimal linear estimation fusion-Part I: Unified fusion rules," *IEEE Trans. Inf. Theory*, vol. 49, no. 9, Sep. 2003.

[5] O. Ledoit and M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *J. Empirical Finance*, 2003, doi: 10.1016/S0927-5398(03)00007-0.

[6] E. J. Candès, "Modern statistical estimation via oracle inequalities," *Acta Numer.*, pp. 169, 2006, doi: 10.1017/S0962492904000.

[7] X. Chen, A. O. HeroIII, and S. Savarese, "Multimodal video indexing and retrieval using directed information," *IEEE Trans. Multimedia*, vol. 14, no. 1, Feb. 2012.

[8] S. M. Kay, *Estimation theory*, Prentice Hall, 1993.

[9] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge, U.K.: Cambridge Univ. Press, 2004.

[10] J. A. Gubner, *Discrete Random Variables*, Cambridge, U.K.: Cambridge Univ. Press, 2006.

[11] J. F. Traub, *Iterative Methods for the Solution of Equations*, Providence, RI, USA: AMS, 1982.