Efficient Combination of Partial Monte Carlo Estimators

David Luengo[⋆], Luca Martino[†], Víctor Elvira[‡], Mónica Bugallo[♯,*]

[⋆] Dep. of Signal Theory and Communic., Universidad Politécnica de Madrid, 28031 Madrid (Spain)

[†] Dep. of Mathematics and Statistics, University of Helsinki, 00014 Helsinki (Finland)

[‡] Dep. of Signal Theory and Communic., Universidad Carlos III de Madrid, 28911 Leganés (Spain)

[♯] Dep. of Electrical and Computer Eng., Stony Brook University, Stony Brook, NY 11794 (USA)

david.luengo@upm.es, lukatotal@gmail.com, velvira@tsc.uc3m.es, monica.bugallo@stonybrook.edu

Abstract

In many practical scenarios, including those dealing with large data sets, calculating global estimators of unknown variables of interest becomes unfeasible. A common solution is obtaining partial estimators and combining them to approximate the global one. In this technical report, we focus on minimum mean squared error (MMSE) estimators, introducing two efficient linear schemes for the fusion of partial estimators. The proposed approaches are valid for any type of partial estimators, although in the simulated scenarios we concentrate on the combination of Monte Carlo estimators due to the nature of the problem addressed. Numerical results show the good performance of the novel fusion methods with only a fraction of the cost of the asymptotically optimal solution.

1 General problem statement

In many applications, we are interested in inferring a variable of interest given a set of observations or measurements. Let us consider the variable of interest, $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^D$, and let $\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^N$ be the observed data. The posterior pdf is then given by

$$p(\mathbf{x}|\mathbf{y}) = \frac{\mathcal{L}(\mathbf{y}|\mathbf{x})g(\mathbf{x})}{Z(\mathbf{y})}, \tag{1}$$

where $\mathcal{L}(\mathbf{y}|\mathbf{x})$ is the likelihood function, $g(\mathbf{x})$ is the prior pdf and $Z(\mathbf{y})$ is the model evidence or partition function. In general, $Z(\mathbf{y})$ is unknown, so we consider the corresponding (usually unnormalized) target pdf,

$$\pi(\mathbf{x}, \mathbf{y}) = \mathcal{L}(\mathbf{y}|\mathbf{x})g(\mathbf{x}), \tag{2}$$

such that $p(\mathbf{x}|\mathbf{y}) = \frac{1}{Z(\mathbf{y})}\pi(\mathbf{x}, \mathbf{y})$, i.e., $p(\mathbf{x}|\mathbf{y}) \propto \pi(\mathbf{x}, \mathbf{y})$.[1]

The problem that we intend to address is computing efficiently some moment of $\mathbf{x}$, i.e., an integral measure w.r.t. the target pdf,

$$I(\mathbf{y}) = \int_{\mathcal{X}} f(\mathbf{x})p(\mathbf{x}|\mathbf{y})d\mathbf{x} = \frac{1}{Z(\mathbf{y})}\int_{\mathcal{X}} f(\mathbf{x})\pi(\mathbf{x}, \mathbf{y})d\mathbf{x}, \tag{3}$$

where the unknown partition function can be obtained as

$$Z(\mathbf{y}) = \int_{\mathcal{X}} \pi(\mathbf{x}, \mathbf{y})d\mathbf{x}. \tag{4}$$

*This work has been supported by the Spanish government's projects COMONSENS (CSD2008-00010), ALCIT (TEC2012-38800-C03-01), AGES (S2010/BMD-2422), DISSECT (TEC2012-38058-C03-01), OTOSiS (TEC2013-41718-R), and COMPREHENSION (TEC2012-38883-C02-01); by the ERC grant 239784 and AoF grant 251170; by the European Union 7th FP through the Marie Curie ITN MLPM2012 (Grant No. 316861); and by the National Science Foundation under Award CCF-0953316.

[1] Note that, for the sake of simplicity and since the observations are fixed, we will often use $\pi(\mathbf{x})$ instead of $\pi(\mathbf{x}, \mathbf{y})$.

However, computing (3) and/or (4) directly is unfeasible in most problems of interest, especially in high dimensional cases (i.e., when $\mathcal{X} \subseteq \mathbb{R}^D$ for a large value of D). In this case, a practical solution is to use a Monte Carlo (MC) algorithm to estimate jointly I and Z. Assuming that we can generate M random samples from the posterior pdf, $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}$ with $\mathbf{x}^{(m)} \sim p(\mathbf{x}|\mathbf{y}) \propto \pi(\mathbf{x}, \mathbf{y})$ for $m = 1, \ldots, M$, the underlying idea is approximating the integral $I(\mathbf{y})$ as

$$\hat{I}(\mathbf{y}) = \frac{1}{M} \sum_{m=1}^{M} f(\mathbf{x}^{(m)}). \tag{5}$$

It can be shown that (5) is an unbiased and consistent estimator of (3), i.e., $\mathbb{E}(\hat{I}) = I$, with $\mathbb{E}(\cdot)$ denoting mathematical expectation, and $\mathrm{Cov}(\hat{I}) = \mathbb{E}((\hat{I} - I)(\hat{I} - I)^\top) \to \mathbf{0}$ as $M \to \infty$. Thus, the problem is developing algorithms to draw samples efficiently from $p(\mathbf{x}|\mathbf{y})$, which is the goal of the so-called *random samplers* [14].

Unfortunately, the estimation of $Z(\mathbf{y})$ is not straightforward in the previous approach, where all the random samples are weighted equally in the computation of $\hat{I}(\mathbf{y})$. An alternative approach, based on *weighted samples*, enables us to estimate jointly both $I(\mathbf{y})$ and $Z(\mathbf{y})$. Let us consider that we draw M random samples from a (normalized) proposal pdf $q(\mathbf{x})$, i.e., $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}$ with $\mathbf{x}^{(m)} \sim q(\mathbf{x})$ for $m = 1, \ldots, M$. Then, we can define the *importance weights*

$$w(\mathbf{x}^{(m)}) = \frac{\pi(\mathbf{x}^{(m)})}{q(\mathbf{x}^{(m)})}, \quad m = 1, \ldots, M, \tag{6}$$

and build the estimators,

$$\hat{Z}(\mathbf{y}) = \frac{1}{M} \sum_{m=1}^{M} w(\mathbf{x}^{(m)}), \tag{7}$$

and

$$\begin{aligned} \hat{I}(\mathbf{y}) &= \frac{1}{\sum_{m=1}^{M} w(\mathbf{x}^{(m)})} \sum_{m=1}^{M} w(\mathbf{x}^{(m)}) f(\mathbf{x}^{(m)}) \\ &= \frac{1}{M\hat{Z}(\mathbf{y})} \sum_{m=1}^{M} w(\mathbf{x}^{(m)}) f(\mathbf{x}^{(m)}). \end{aligned} \tag{8}$$

On the one hand, it can be shown that (7) is an unbiased and consistent estimator of $Z(\mathbf{y})$. On the other hand, (8) is a consistent but asymptotically biased estimator of $I(\mathbf{y})$ (i.e., it is biased due to the presence of $Z(\mathbf{y})$ in the denominator, but the bias tends to zero as $M \to \infty$).

In this technical report we concentrate on the linear combination of partial Monte Carlo (MC) estimators, developing efficient fusion rules that can be applied both for MCMC-based and importance sampling estimators. Indeed, the rules developed for any class of unbiased partial estimators, though we focus on MC-based estimators since they are often the only alternative for performing optimal estimation in complicated real-world problems. Table 1 summarizes the notation used throughout the remainder of the technical report. The report is organized as follows. First of all, Section 2 discusses the problem of global vs. partial estimation. This section addresses the optimality of linear fusion rules (which will be the focus of the rest of the report), an issue that is illustrated by a simple example in Appendix A. Then, in Section 3 we derive the optimal linear combination of partial estimators from a Bayesian perspective. Section 4 is devoted to obtaining both the optimal and efficient sub-optimal estimators based on a constrained optimization approach. Related works are then briefly reviewed in Section 5. Finally, Section 6 shows some preliminary numerical simulations and Section 7 provides the conclusions and some potential future lines.

Table 1: Summary of the Notation.

$\mathbf{x}$	Variable of interest to be estimated.
D	Dimension of the variable of interest $\mathbf{x}$.
$\mathbf{y}$	Vector of observations.
N	Number of observations (i.e., dimension of $\mathbf{y}$).
M	Total number of particles.
L	Number of parallel (partial) estimators.
N_ℓ, M_ℓ	Number of data/particles for the ℓ-th estimator ($1 \leq \ell \leq L$).
$\mathbf{y}_\ell$	Data set for the ℓ-th estimator ($1 \leq \ell \leq L$).
$p(\mathbf{x}\|\mathbf{y})$	Global posterior pdf.
$p_\ell(\mathbf{x}\|\mathbf{y}_\ell)$	Partial posterior pdf for the ℓ-th estimator ($1 \leq \ell \leq L$).
$\pi(\mathbf{x}, \mathbf{y})$	Global target pdf.
$\pi_\ell(\mathbf{x}, \mathbf{y}_\ell)$	Partial target pdf for the ℓ-th estimator ($1 \leq \ell \leq L$).
$Z(\mathbf{y})$	Global partition function.
$Z_\ell(\mathbf{y}_\ell)$	Partial partition function for the ℓ-th estimator ($1 \leq \ell \leq L$).

2 Inference for Big Data: Global vs. Partial Estimators

Let us assume a fixed model, so that the likelihood and the priors are fixed. The minimum mean squared error (MMSE) estimator of $\mathbf{x}$ is then given by (3) using $f(\mathbf{x}) = \mathbf{x}$:[2]

$$\hat{\mathbf{x}}^{(\text{MMSE})} = I(\mathbf{y}) = \mathbb{E}(\mathbf{x}|\mathbf{y}) = \int_{\mathcal{X}} \mathbf{x}\ p(\mathbf{x}|\mathbf{y}) d\mathbf{x} = \frac{1}{Z(\mathbf{y})} \int_{\mathcal{X}} \mathbf{x}\ \pi(\mathbf{x}, \mathbf{y}) d\mathbf{x}, \tag{9}$$

Now, let us assume that we are facing a big data problem, where we cannot deal with the whole data set globally.[3] Then, a natural solution is splitting the data in L groups/clusters, so that the ℓ-th cluster ($1 \leq \ell \leq L$) only has access to N_ℓ samples. For the sake of simplicity, let us assume that the samples have been assigned randomly to each cluster and that each sample can only belong to one cluster, i.e., we have disjoint sets of samples such that $N = \sum_{\ell=1}^{L} N_\ell$. Then, we can obtain the partial MMSE estimator for each cluster as[4]

$$\hat{\mathbf{x}}_\ell^{(\text{MMSE})} = I_\ell(\mathbf{y}_\ell) = \mathbb{E}(\mathbf{x}|\mathbf{y}_\ell) = \int_{\mathcal{X}} \mathbf{x}\ p_\ell(\mathbf{x}|\mathbf{y}_\ell) d\mathbf{x} = \frac{1}{Z_\ell(\mathbf{y}_\ell)} \int_{\mathcal{X}} \mathbf{x}\ \pi_\ell(\mathbf{x}, \mathbf{y}_\ell) d\mathbf{x}, \tag{10}$$

Our goal now is obtaining the global MMSE estimator, $\hat{\mathbf{x}}^{(\text{MMSE})}$, from the set of partial MMSE estimators, $\hat{\mathbf{x}}_\ell^{(\text{MMSE})}$ for $\ell = 1,\ \ldots,\ L$.[5]

In this technical report we consider only the communications-free situation for the partial estimators, i.e., we assume that the partial estimators can only transmit their final estimators to the fusion center (FC) and are not allowed to communicate with each other during the estimation process. The FC will then be the responsible for combining all the estimates in an efficient way to obtain the global MMSE estimator (if it is feasible) or at least the best possible approximation. With respect to this goal, let us remark that the exact global MMSE estimator cannot be attained in general from the partial MMSE estimators. Indeed, the only situation where the exact global MMSE estimator can be attained from the partial MMSE estimators is when the problem is "separable in the data". In this situation, the exact global MMSE estimator can be built as a (possibly nonlinear) function of the partial MMSE estimators:

$$\hat{\mathbf{x}}^{(\text{MMSE})} = h\left(\hat{\mathbf{x}}_1^{(\text{MMSE})},\ \ldots,\ \hat{\mathbf{x}}_L^{(\text{MMSE})}\right). \tag{11}$$

[2]Note that the MMSE estimator of $\boldsymbol{\theta} = f(\mathbf{x})$ would be given exactly by (3).

[3]Even when we can deal with the whole data set globally, splitting it into L data sets may be a good idea. This is due to the fact that the posterior pdf tends to become more "peaky" as the number of data increases, thus making inferences harder, especially for complex multi-modal problems.

[4]Note that this is the MMSE estimator of $\mathbf{x}$ given all the data directly available to the ℓ-th estimator, i.e., $\mathbf{y}_\ell$.

[5]Note that we prefer to use the name *partial* MMSE estimator instead of *local* MMSE estimator to emphasize the fact that $\hat{\mathbf{x}}_\ell^{(\text{MMSE})}$ corresponds to the MMSE estimator of the complete set of variables of interest obtained using only partial information.

A particular case of this situation occurs when the global MMSE estimator is a weighted linear combination of the partial MMSE estimators:

$$\hat{\mathbf{x}}^{(\mathrm{MMSE})} = \sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell \hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})}, \tag{12}$$

where $\mathbf{\Lambda}_\ell$ is a $D \times D$ weighting matrix. Note that this is equivalent to stating that

$$I(\mathbf{y}) = \int_{\mathcal{X}} \mathbf{x}\; p(\mathbf{x}|\mathbf{y}) d\mathbf{x} = \sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell \int_{\mathcal{X}} \mathbf{x}\; p_\ell(\mathbf{x}|\mathbf{y}_\ell) d\mathbf{x} = \sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell I_\ell(\mathbf{y}_\ell) \tag{13}$$

For instance, this happens when the posterior pdf is Gaussian with a mean that is a weighted linear combination of the data (see Appendix A for a simple example of this situation). In this case, the partial estimators may work independently (i.e., without any communication until the final partial values have been obtained) and the main concern is obtaining the optimal nonlinear combination $h(\cdot)$ or the weighting matrices $\mathbf{\Lambda}_\ell$ in the linear case. If the problem is not "separable in the data", a linear combination (or a non-linear one in the most general case) of the partial MMSE estimators can still be constructed, but in this case we will only obtain an approximation of the global MMSE estimator and not the exact global MMSE estimator. Nevertheless, in the following section we show that the linear combination of partial MMSE estimators is asymptotically optimal, i.e., as the number of data increases the exact global MMSE estimator is given precisely by the linear combination of partial MMSE estimators.

3 Optimal Linear Combination of Partial MMSE Estimators: A Bayesian Perspective

3.1 Derivation based on the Bernstein-von Mises Theorem

In general, the MMSE estimator is a non-linear function of the whole data set and the exact global MMSE estimator cannot be attained by any combination of partial MMSE estimators.[6] However, the Bernstein-von Mises (a.k.a. Bayesian central limit) theorem states that, under suitable regularity conditions, the partial posterior PDFs, $p_\ell(\mathbf{x}|\mathbf{y}_\ell)$, converge to Gaussian PDFs as N_ℓ tends to infinity [8, 16], i.e.,

$$p_\ell(\mathbf{x}|\mathbf{y}_\ell) \to \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\mathbf{x}}^{(\ell)}, \mathbf{C}_{\mathbf{x}}^{(\ell)}) \qquad \text{as} \quad N_\ell \to \infty, \tag{14}$$

with $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\mathbf{x}}^{(\ell)}, \mathbf{C}_{\mathbf{x}}^{(\ell)})$ indicating that $\mathbf{x}$ has a Gaussian PDF with a mean vector $\boldsymbol{\mu}_{\mathbf{x}}^{(\ell)} = \hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})}$ and a covariance matrix

$$\begin{aligned} \mathbf{C}_{\mathbf{x}}^{(\ell)} &= \mathbb{E}\left((\hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})} - \mathbf{x})(\hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})} - \mathbf{x})^\top \right) \\ &= \int_{\mathcal{X}} (\hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})} - \mathbf{x})(\hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})} - \mathbf{x})^\top p_\ell(\mathbf{x}|\mathbf{y}_\ell) d\mathbf{x}. \end{aligned} \tag{15}$$

Assuming that we have independent (though not necessarily identically distributed) observations and that each of them can only belong to one cluster (i.e., we have disjoint sets of samples such that $N = \sum_{\ell=1}^{L} N_\ell$), the global posterior PDF also converges to a Gaussian PDF as N tends to infinity, i.e.,

$$p(\mathbf{x}|\mathbf{y}) = \prod_{\ell=1}^{L} p_\ell(\mathbf{x}|\mathbf{y}_\ell) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\mathbf{x}}, \mathbf{C}_{\mathbf{x}}) \qquad \text{as} \quad N \to \infty, \tag{16}$$

[6] An exception occurs when the global MMSE estimator is "separable in the data". For instance, this happens when the global posterior PDF is Gaussian with a mean that is a weighted linear combination of the data. In this case, a properly weighted linear combination of the partial MMSE estimators leads to the exact global MMSE estimator.

with

$$\mathbf{C}_{\mathbf{x}} = \left[\sum_{\ell=1}^{L} \left(\mathbf{C}_{\mathbf{x}}^{(\ell)}\right)^{-1}\right]^{-1}, \tag{17a}$$

$$\boldsymbol{\mu}_{\mathbf{x}} = \mathbf{C}_{\mathbf{x}} \sum_{\ell=1}^{L} \left(\mathbf{C}_{\mathbf{x}}^{(\ell)}\right)^{-1} \hat{\mathbf{x}}_{\ell}^{(\mathrm{MMSE})}. \tag{17b}$$

Eq. (17b) has been already proposed in [15] as the optimal linear combination of communication-free distributed Monte Carlo estimators, and has also been exploited in [11] to obtain asymptotically exact samples from the global posterior by sampling from a multivariate Gaussian whose covariance matrix and mean vector are given respectively by (17a) and (17b).

3.2 Alternative Derivation based on the Gaussian Distribution of the Partial MMSE Estimators

In this section we show an alternative derivation of the previous result that provides us with a complementary view of the optimal weight choice. From a Bayesian point of view, let us define the *Bayesian risk* as

$$\begin{aligned} R(\hat{\mathbf{x}}) &= \int_{\mathcal{Y}} \int_{\mathcal{X}} C(\mathbf{x}, \hat{\mathbf{x}}(\mathbf{y})) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}, \\ &= \int_{\mathcal{Y}} \int_{\mathcal{X}} C(\mathbf{x}, \hat{\mathbf{x}}(\mathbf{y})) p(\mathbf{x}|\mathbf{y}) p(\mathbf{y}) d\mathbf{x} d\mathbf{y}, \\ &= \int_{\mathcal{Y}} \left[\int_{\mathcal{X}} C(\mathbf{x}, \hat{\mathbf{x}}(\mathbf{y})) p(\mathbf{x}|\mathbf{y}) d\mathbf{x}\right] p(\mathbf{y}) d\mathbf{y} = \int_{\mathcal{Y}} r(\hat{\mathbf{x}}) p(\mathbf{y}) d\mathbf{y}, \end{aligned} \tag{18}$$

where $\hat{\mathbf{x}} = \hat{\mathbf{x}}(\mathbf{y})$ can be any estimator of $\mathbf{x}$,

$$r(\hat{\mathbf{x}}) = \int_{\mathcal{X}} C(\mathbf{x}, \hat{\mathbf{x}}) p(\mathbf{x}|\mathbf{y}) d\mathbf{x}, \tag{19}$$

and $C(\mathbf{x}, \hat{\mathbf{x}})$ is some suitable *cost function*. Since $p(\mathbf{y})$ is a fixed non-negative function (as the observations are fixed and $p(\mathbf{y})$ is a pdf), minimizing (18) is equivalent to minimizing (19). Now, let us consider the quadratic cost,

$$C(\mathbf{x}, \hat{\mathbf{x}}) = (\hat{\mathbf{x}} - \mathbf{x})^{\top} (\hat{\mathbf{x}} - \mathbf{x}), \tag{20}$$

which is the usual cost function for regression problems. Then, (19) becomes[7]

$$r(\hat{\mathbf{x}}) = \mathrm{MSE}(\hat{\mathbf{x}}|\mathbf{y}) = \int_{\mathcal{X}} (\hat{\mathbf{x}} - \mathbf{x})^{\top} (\hat{\mathbf{x}} - \mathbf{x}) p(\mathbf{x}|\mathbf{y}) d\mathbf{x}, \tag{22}$$

and the optimal estimator corresponds to the MMSE estimator, which is given by Eq. (9):

$$\hat{\mathbf{x}}^{(\mathrm{MMSE})} = \mathbb{E}(\mathbf{x}|\mathbf{y}) = \int_{\mathcal{X}} \mathbf{x}\, p(\mathbf{x}|\mathbf{y}) d\mathbf{x} = \frac{1}{Z(\mathbf{y})} \int_{\mathcal{X}} \mathbf{x}\, \pi(\mathbf{x}, \mathbf{y}) d\mathbf{x}.$$

[7] In the Bayesian literature, $r(\hat{\mathbf{x}})$, as given by (22), is usually known as *Bayesian Expected Loss*. The *Bayesian MSE* is obtained performing a double integral on both the data and the parameters of interest using the joint pdf $p(\mathbf{x}, \mathbf{y})$ (i.e., inserting the quadratic loss function of (20) in (18)):

$$\mathrm{MSE}(\hat{\mathbf{x}}) = \int_{\mathcal{Y}} \int_{\mathcal{X}} (\hat{\mathbf{x}} - \mathbf{x})^{\top} (\hat{\mathbf{x}} - \mathbf{x}) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}. \tag{21}$$

Hence, strictly speaking Eq. (22) does not correspond to the Bayesian MSE. However, by assuming that the data are fixed we can remove the outer integral in (21) and perform the integration only on $\mathbf{x}$ using $p(\mathbf{x}|\mathbf{y})$. In order to distinguish this *conditional MSE* from the *full Bayesian MSE* we use the notation $\mathrm{MSE}(\hat{\mathbf{x}}|\mathbf{y})$ instead of simply $\mathrm{MSE}(\hat{\mathbf{x}})$. However, for the sake of simplicity, in the following we will refer to it just as the MSE. Thus, whenever we mention the MSE in the sequel we will be referring to the conditional MSE as defined by (22).

Let us consider that our observations are the outputs of each of the L partial MMSE estimators, $\hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})}$, which are independent and follow Gaussian distributions with means equal to the true parameter vector $\mathbf{x}$ and covariance matrices $\mathbf{C}_\mathbf{x}^{(\ell)}$.[8] Then, $\pi(\mathbf{x})$ is given by

$$\begin{aligned}\pi(\mathbf{x}) &= \prod_{\ell=1}^{L} \mathcal{N}(\hat{\mathbf{x}}_\ell|\mathbf{x}, \mathbf{C}_\mathbf{x}^{(\ell)}) \\ &= \prod_{\ell=1}^{L} (2\pi)^{-D/2}|\mathbf{C}_\mathbf{x}^{(\ell)}|^{-1/2} \exp\left(-\frac{1}{2}(\hat{\mathbf{x}}_\ell - \mathbf{x})^\top \left(\mathbf{C}_\mathbf{x}^{(\ell)}\right)^{-1} (\hat{\mathbf{x}}_\ell - \mathbf{x})\right) \\ &= \prod_{\ell=1}^{L} (2\pi)^{-D/2}|\mathbf{C}_\mathbf{x}^{(\ell)}|^{-1/2} \times \exp\left(-\frac{1}{2}\sum_{\ell=1}^{L}(\hat{\mathbf{x}}_\ell - \mathbf{x})^\top \left(\mathbf{C}_\mathbf{x}^{(\ell)}\right)^{-1} (\hat{\mathbf{x}}_\ell - \mathbf{x})\right) \end{aligned} \tag{23}$$

It is straightforward to see that

$$\pi(\mathbf{x}) \propto (2\pi)^{-D/2}|\mathbf{C}_\mathbf{x}|^{-1/2} \times \exp\left(-\frac{1}{2}(\hat{\mathbf{x}} - \boldsymbol{\mu}_\mathbf{x})^\top \mathbf{C}_\mathbf{x}^{-1}(\hat{\mathbf{x}} - \boldsymbol{\mu}_\mathbf{x})\right), \tag{24}$$

where

$$\begin{aligned}\mathbf{C}_\mathbf{x} &= \left[\sum_{\ell=1}^{L}\left(\mathbf{C}_\mathbf{x}^{(\ell)}\right)^{-1}\right]^{-1}, \\ \boldsymbol{\mu}_\mathbf{x} &= \mathbf{C}_\mathbf{x}\sum_{\ell=1}^{L}\left(\mathbf{C}_\mathbf{x}^{(\ell)}\right)^{-1}\hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})}.\end{aligned} \tag{25}$$

Hence, the global MMSE estimator is finally given by

$$\hat{\mathbf{x}}^{(\mathrm{MMSE})} = \sum_{\ell=1}^{L}\left[\sum_{k=1}^{L}\left(\mathbf{C}_\mathbf{x}^{(k)}\right)^{-1}\right]^{-1}\left(\mathbf{C}_\mathbf{x}^{(\ell)}\right)^{-1}\hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})}. \tag{26}$$

3.3 Particular Cases

Note that Eq. (26) requires computing a $D \times D$ weight matrix for each of the L estimators:

$$\boldsymbol{\Lambda}_\ell = \left[\sum_{k=1}^{L}\left(\mathbf{C}_\mathbf{x}^{(k)}\right)^{-1}\right]^{-1}\left(\mathbf{C}_\mathbf{x}^{(\ell)}\right)^{-1}. \tag{27}$$

This implies computing up to D^2L weights overall, which may be unfeasible (or at least very costly from a computational/storage point of view) when D and/or L is large, and is due to two facts:

1. The parameters to be estimated are interrelated (i.e., the $\mathbf{C}_\mathbf{x}^{(\ell)}$ are non-diagonal matrices).

2. The quality of each of the partial estimators is different (i.e., the $\mathbf{C}_\mathbf{x}^{(\ell)}$ are different).

However, in certain cases the optimum weight matrix may contain a reduced number of coefficients. Furthermore, "reduced matrices" can always be used to obtain an approximation of the optimal case.

Let us consider first the case where the parameters are not interrelated. Then, the covariance matrix for the ℓ-th estimator will be given by

$$\mathbf{C}_\mathbf{x}^{(\ell)} = \mathrm{diag}\left(\sigma_{\ell,1}^2,\ \ldots,\ \sigma_{\ell,D}^2\right), \tag{28}$$

[8] When we have a large number of data available for each partial estimator, due to the central limit theorem it is reasonable to assume that each of these estimators follows a Gaussian distribution.

with

$$\sigma_{\ell,d}^2 = \int_{\mathcal{X}_d} (\hat{x}_{\ell,d} - x_d)^2 p(x_d|\mathbf{y}_\ell) dx_d, \tag{29}$$

and the optimal weight matrix will be

$$\mathbf{\Lambda}_\ell = \text{diag}\left(\alpha_{\ell,1}^2,\ \ldots,\ \alpha_{\ell,D}^2\right), \tag{30}$$

with

$$\alpha_{\ell,d} = \frac{\sigma_{\ell,d}^{-2}}{\sum_{k=1}^{L} \sigma_{k,d}^{-2}}. \tag{31}$$

In this case, only D parameters are required for each of the L estimators (i.e., DL parameters overall). If we want to reduce the number of parameters further, then we can consider using a single parameter per estimator (i.e., only L parameters overall) obtained averaging (31) over the set of all the parameters:

$$\alpha_\ell = \frac{1}{D}\sum_{d=1}^{D} \alpha_{\ell,d} = \frac{1}{D}\sum_{d=1}^{D} \left(\sum_{k=1}^{L} \sigma_{k,d}^{-2}\right)^{-1} \sigma_{\ell,d}^{-2}. \tag{32}$$

When the partial estimators have the same variance for all the parameters (i.e., $\sigma_{\ell,d}^{-2} = \sigma_\ell^{-2}$ for $1 \leq d \leq D$, and they are equally hard to estimate), then (32) becomes

$$\alpha_\ell = \frac{\sigma_\ell^{-2}}{\sum_{k=1}^{L} \sigma_k^{-2}}, \tag{33}$$

and corresponds to the optimal weights in this case. Finally, when all the partial estimators have the same quality (regardless of whether the parameters are related or not) we have simply $\alpha_\ell = 1/L$.

4 Efficient Linear Fusion of Partial MMSE Estimators: A Constrained Optimization Approach

Let us assume that we have a problem which is "separable in the data" in such a way that

$$\hat{\mathbf{x}}^{(\text{MMSE})} = \sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell \hat{\mathbf{x}}_\ell^{(\text{MMSE})}, \tag{34}$$

with $\hat{\mathbf{x}}_\ell^{(\text{MMSE})} = \hat{\mathbf{x}}_\ell$ being the true MMSE estimator given by (10). However, instead of the true partial MMSE estimators we can only compute appropriate "local" approximations (e.g., based on an MCMC method or importance sampling). Let us also assume that $\mathbb{E}(\hat{\mathbf{x}}_\ell) = \mathbf{x}$ (i.e., the partial estimators are unbiased) and

$$\text{Cov}(\hat{\mathbf{x}}_\ell) = \mathbb{E}\left((\hat{\mathbf{x}}_\ell - \mathbf{x})(\hat{\mathbf{x}}_\ell - \mathbf{x})^\top\right) = \mathbf{C}_{\mathbf{x}}^{(\ell)},$$

with $\text{Cov}(\hat{\mathbf{x}}_\ell) \to \mathbf{0}$ as $M_\ell \to \infty$ (i.e., the partial estimators are consistent). The issue now is how to obtain the optimal collection of weights α_ℓ $(1 \leq \ell \leq L)$ such that

$$\hat{\mathbf{x}} = \sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell \hat{\mathbf{x}}_\ell = \hat{\mathbf{x}}^{(\text{MMSE})}. \tag{35}$$

Note that (35) will not be the global MMSE estimator when the problem is not "separable in the data". However, it will provide us with the best approximation that we can obtain by using a linear combination of the partial MMSE estimators and will become optimal asymptotically when $N \to \infty$, as shown in Sections 3.1 and 3.2. In the sequel we derive the optimum coefficients for the linear combination following a constrained minimization approach. We start by considering two particular cases (a single coefficient per estimator and a diagonal matrix $\mathbf{\Lambda}_\ell$) and then we finally address the most general case.

4.1 Particular Case: Single Coefficient per Estimator

As mentioned before, this problem may be approached by trying to minimize the mean squared error (MSE) of the global estimator without any further assumptions on the partial estimators. Let us consider the particular case in which a single coefficient per estimator is used to construct the global estimator:

$$\hat{\mathbf{x}} = \sum_{\ell=1}^{L} \alpha_\ell \hat{\mathbf{x}}_\ell, \tag{36}$$

which is obtained by setting $\mathbf{\Lambda}_\ell = \alpha_\ell \mathbf{I}$ in (35). Clearly this will provide a suboptimal solution in general, but is a fast and low-cost solution for the combination of estimators, and we can obtain the optimal weights in closed form.

On the one hand, since the partial estimators are unbiased, it is straightforward to see that the mean of the global estimator given by (36) is

$$\mathbb{E}(\hat{\mathbf{x}}) = \sum_{\ell=1}^{L} \alpha_\ell \mathbb{E}(\hat{\mathbf{x}}_\ell) = \left(\sum_{\ell=1}^{L} \alpha_\ell \right) \mathbf{x}. \tag{37}$$

Hence, in order to obtain an unbiased global estimator we need to have

$$\sum_{\ell=1}^{L} \alpha_\ell = 1, \tag{38}$$

which is the first condition that must be fulfilled by the coefficients. On the other hand, the covariance matrix for the global estimator is given by

$$\begin{aligned}
\mathbf{C}_\mathbf{x} = \mathrm{Cov}(\hat{\mathbf{x}}) &= \mathbb{E}\left((\hat{\mathbf{x}} - \mathbb{E}(\hat{\mathbf{x}}))(\hat{\mathbf{x}} - \mathbb{E}(\hat{\mathbf{x}}))^\top \right) \\
&= \mathbb{E}\left(\sum_{\ell=1}^{L} \alpha_\ell (\hat{\mathbf{x}}_\ell - \mathbb{E}(\hat{\mathbf{x}}_\ell)) \sum_{k=1}^{L} \alpha_k (\hat{\mathbf{x}}_k - \mathbb{E}(\hat{\mathbf{x}}_k))^\top \right) \\
&= \sum_{\ell=1}^{L} \sum_{k=1}^{L} \alpha_\ell \alpha_k \mathbb{E}\left((\hat{\mathbf{x}}_\ell - \mathbb{E}(\hat{\mathbf{x}}_\ell))(\hat{\mathbf{x}}_k - \mathbb{E}(\hat{\mathbf{x}}_k))^\top \right) \\
&= \sum_{\ell=1}^{L} \alpha_\ell^2 \mathbb{E}\left((\hat{\mathbf{x}}_\ell - \mathbb{E}(\hat{\mathbf{x}}_\ell))(\hat{\mathbf{x}}_\ell - \mathbb{E}(\hat{\mathbf{x}}_\ell))^\top \right) = \sum_{\ell=1}^{L} \alpha_\ell^2 \mathbf{C}_\mathbf{x}^{(\ell)},
\end{aligned} \tag{39}$$

where the last expression comes from the assumption that the partial estimators are independent (i.e., $\mathbb{E}\left((\hat{\mathbf{x}}_\ell - \mathbb{E}(\hat{\mathbf{x}}_\ell))(\hat{\mathbf{x}}_k - \mathbb{E}(\hat{\mathbf{x}}_k))^\top \right) = \mathbf{C}_\mathbf{x}^{(\ell)} \delta[\ell - k]$ with $\delta[\ell - k]$ denoting Kronecker's delta).[9] Now, let us remark that the MSE can be expressed as a function of the global covariance matrix:

$$\mathrm{MSE}(\hat{\mathbf{x}}|\mathbf{y}) = \mathbb{E}\left((\hat{\mathbf{x}} - \mathbf{x})^\top (\hat{\mathbf{x}} - \mathbf{x}) \right) = \mathbb{E}\left(\mathrm{Tr}\left((\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^\top \right) \right) = \mathrm{Tr}\left(\mathbb{E}\left((\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^\top \right) \right) = \mathrm{Tr}(\mathbf{C}_\mathbf{x}), \tag{40}$$

where $\mathrm{Tr}(\mathbf{C}_\mathbf{x})$ denotes the trace of the global covariance matrix:

$$\mathrm{Tr}(\mathbf{C}_\mathbf{x}) = \sum_{d=1}^{D} \mathbf{C}_\mathbf{x}[d,d] = \sum_{d=1}^{D} \sigma_{x_d}^2, \tag{41}$$

with $\sigma_{x_d}^2 = \mathbb{E}((\hat{x}_d - x_d)^2)$. Substituting the last expression of (39) in (40), the MSE is finally given by

$$\mathrm{MSE}(\hat{\mathbf{x}}|\mathbf{y}) = \mathrm{Tr}\left(\sum_{\ell=1}^{L} \alpha_\ell^2 \mathbf{C}_\mathbf{x}^{(\ell)} \right) = \sum_{\ell=1}^{L} \alpha_\ell^2 \mathrm{Tr}\left(\mathbf{C}_\mathbf{x}^{(\ell)} \right), \tag{42}$$

[9] Note that combining non-independent estimators is a much harder problem and it may even be pointless if the dependence is large, as the new information incorporated by each estimator may be negligibly small. However, we plan to consider this issue in future works, as well as the combination of biased estimators.

where $\mathrm{Tr}(\mathbf{C}_{\mathbf{x}}^{(\ell)})$ denotes the trace of the ℓ-th partial covariance matrix:

$$\mathrm{Tr}(\mathbf{C}_{\mathbf{x}}^{(\ell)}) = \sum_{d=1}^{D} \mathbf{C}_{\mathbf{x}}^{(\ell)}[d,d] = \sum_{d=1}^{D} \sigma_{\ell,d}^2, \tag{43}$$

with $\sigma_{\ell,d}^2 = \mathbb{E}((\hat{x}_d^{(\ell)} - x_d)^2)$.

Our goal now is finding the set of α_ℓ that minimize (42), subject to Eq. (38) in order to obtain an unbiased estimator. From the previous equations we can see that the optimal selection of the weights can be formulated as a constrained optimization problem:

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \sum_{\ell=1}^{L} \alpha_\ell^2 \mathrm{Tr}\left(\mathbf{C}_{\mathbf{x}}^{(\ell)}\right), \tag{44a}$$

$$\text{s.t.} \qquad \sum_{\ell=1}^{L} \alpha_\ell = 1, \tag{44b}$$

with $\boldsymbol{\alpha} = [\alpha_1,\ \ldots,\ \alpha_L]^\top$. Eqs. (44a) and (44b) corresponds to a convex optimization problem: (44a) corresponds to a standard MSE minimization problem and (44b) is required to guarantee that the resulting global estimator is unbiased. The convexity ensures that the solution of this problem is unique and, since the only restriction in (44b) has to be fulfilled with equality, it can be reformulated as an unconstrained optimization problem using the method of Lagrange multipliers [3]:

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} J(\boldsymbol{\alpha}), \tag{45}$$

where

$$J(\boldsymbol{\alpha}) = \sum_{\ell=1}^{L} \alpha_\ell^2 \mathrm{Tr}(\mathbf{C}_{\mathbf{x}}^{(\ell)}) + \lambda \left(\sum_{\ell=1}^{L} \alpha_\ell - 1 \right), \tag{46}$$

and λ is the Lagrange multiplier associated to the problem. Differentiating (46) w.r.t. α_ℓ and λ we obtain a set of $L+1$ linear equations:

$$\begin{aligned} 2\alpha_\ell \mathrm{Tr}(\mathbf{C}_{\mathbf{x}}^{(\ell)}) + \lambda = 0, \qquad & \text{for} \quad \ell = 1,\ \ldots,\ L; \\ \sum_{\ell=1}^{L} \alpha_\ell = 1. & \end{aligned} \tag{47}$$

In matrix form this system of equations may be expressed more compactly as

$$\begin{bmatrix} \mathbf{A} & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}, \tag{48}$$

where $\mathbf{1}$ indicates an $L \times 1$ vector of ones and $\mathbf{A} = \mathrm{diag}(2T_1,\ \ldots,\ 2T_L)$ is a diagonal matrix, with $T_\ell = \mathrm{Tr}(\mathbf{C}_{\mathbf{x}}^{(\ell)})$. The solution of (48) can be expressed as

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{P} & \mathbf{q} \\ \mathbf{q}^\top & m \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{q} \\ m \end{bmatrix}, \tag{49}$$

where the required terms in (49) ($\mathbf{P}$, $\mathbf{q}$ and m) can be found using the matrix inversion lemma:

$$\begin{aligned}
\mathbf{P} &= \mathbf{A}^{-1}[\mathbf{I} + m\mathbf{1}\mathbf{1}^\top\mathbf{A}^{-1}], \\
\mathbf{A}^{-1} &= \operatorname{diag}\left(\frac{1}{2}T_1^{-1},\ \ldots,\ \frac{1}{2}T_L^{-1}\right), \\
\mathbf{q} &= [q_1,\ \ldots,\ q_L]^\top, \\
q_\ell = \alpha_\ell &= \frac{T_\ell^{-1}}{\sum_{k=1}^{L} T_k^{-1}}, \qquad \text{for} \quad \ell = 1,\ \ldots,\ L, \\
m = \lambda &= -\frac{2}{\sum_{k=1}^{L} T_k^{-1}}.
\end{aligned} \tag{50}$$

Hence, the global single coefficient MMSE (SCMSE) estimator for the separable case is given by

$$\hat{\mathbf{x}}^{(\mathrm{SCMSE})} = \sum_{\ell=1}^{L} \frac{T_\ell^{-1}}{\sum_{k=1}^{L} T_k^{-1}} \hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})} = \sum_{\ell=1}^{L} \frac{\left[\mathrm{MSE}(\hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})}|\mathbf{y}_\ell)\right]^{-1}}{\sum_{k=1}^{L}\left[\mathrm{MSE}(\hat{\mathbf{x}}_k^{(\mathrm{MMSE})}|\mathbf{y}_k)\right]^{-1}} \hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})}. \tag{51}$$

For the non-separable case (51) will be different from the global MMSE estimator, but it will be the best linear unbiased estimator (in terms of minimizing the MSE) that can be constructed from the local MMSE estimators without any further assumption using a single coefficient.

4.2 Particular Case: Diagonal Weighting Matrices

The SCMSE estimator has a substantially reduced computational cost w.r.t. the LMSE estimator, since it only requires the estimation of L parameters overall instead of the D^2L parameters of the LMSE estimator. However, noting that the optimal weights in (51) involve the trace of the partial covariance matrices, we introduce an independent linear minimum mean squared estimator (ILMSE) where $\mathbf{\Lambda}_\ell = \operatorname{diag}(\alpha_{\ell,1},\ \ldots,\ \alpha_{\ell,D})$. This approach leads to an independent estimation of each of the D variables of interest:

$$\hat{x}_d^{(\mathrm{ILMSE})} = \sum_{\ell=1}^{L} \alpha_{\ell,d}\ \hat{x}_{\ell,d}^{(\mathrm{MMSE})}, \tag{52}$$

where $1 \leq d \leq D$ and $\hat{x}_{\ell,d}^{(\mathrm{MMSE})}$ denotes the d-th component of the ℓ-th partial MMSE estimator. In practice, the weights in (52) can be obtained by solving D single parameter constrained optimization problems:

$$\boldsymbol{\alpha}_d^* = \arg\min_{\boldsymbol{\alpha}_d} \sum_{\ell=1}^{L} \alpha_{\ell,d}^2 C_{x_d}^{(\ell)}, \tag{53a}$$

$$\text{s.t.} \qquad \sum_{\ell=1}^{L} \alpha_{\ell,d} = 1, \tag{53b}$$

where $\boldsymbol{\alpha}_d = [\alpha_{1,d},\ \ldots,\ \alpha_{L,d}]^\top$ and $C_{x_d}^{(\ell)}$ is the d-th element along the main diagonal of $\mathbf{C}_{\mathbf{x}}^{(\ell)}$. The solution is now

$$\alpha_{\ell,d} = \frac{\left[\mathrm{MSE}(\hat{x}_{\ell,d}^{(\mathrm{MMSE})}|\mathbf{y}_\ell)\right]^{-1}}{\sum_{k=1}^{L}\left[\mathrm{MSE}(\hat{x}_{k,d}^{(\mathrm{MMSE})}|\mathbf{y}_k)\right]^{-1}}. \tag{54}$$

This approach requires the estimation of DL parameters overall, and thus it can be seen as an intermediate approach between the LMSE and the SCMSE described in the following section.

4.3 General Case: Optimal Linear Combination

Finally, we can also consider the most general combination of estimators, given by Eq. (35):

$$\hat{\mathbf{x}} = \sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell \hat{\mathbf{x}}_\ell.$$

In this case, the mean of the global estimator is given by

$$\mathbb{E}(\hat{\mathbf{x}}) = \sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell \mathbb{E}(\hat{\mathbf{x}}_\ell) = \left(\sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell \right) \mathbf{x}, \tag{55}$$

and the condition required to obtain an unbiased global estimator becomes

$$\sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell = \mathbf{I}. \tag{56}$$

Following a similar procedure as in the previous section, we can also obtain the covariance matrix of the global estimator,

$$\mathbf{C}_\mathbf{x} = \sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell \mathbf{C}_\mathbf{x}^{(\ell)} \mathbf{\Lambda}_\ell^\top, \tag{57}$$

and the MSE,

$$\mathrm{MSE}(\hat{\mathbf{x}}|\mathbf{y}) = \sum_{\ell=1}^{L} \mathrm{Tr}\left(\mathbf{\Lambda}_\ell \mathbf{C}_\mathbf{x}^{(\ell)} \mathbf{\Lambda}_\ell^\top \right). \tag{58}$$

Now we can formulate the generalized version of the constrained optimization problem solved in the previous two sections:

$$\mathbf{\Lambda}^* = \arg\min_{\mathbf{\Lambda}} \sum_{\ell=1}^{L} \mathrm{Tr}\left(\mathbf{\Lambda}_\ell \mathbf{C}_\mathbf{x}^{(\ell)} \mathbf{\Lambda}_\ell^\top \right), \tag{59a}$$

$$\text{s.t.} \qquad \sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell = \mathbf{I}, \tag{59b}$$

with $\mathbf{\Lambda} = [\mathbf{\Lambda}_1, \ \ldots, \ \mathbf{\Lambda}_L]^\top$. Once more, applying the method of Lagrange multipliers we obtain the solution for each of the weight matrices as

$$\mathbf{\Lambda}_\ell = \left[\sum_{k=1}^{L} \left(\mathbf{C}_\mathbf{x}^{(k)} \right)^{-1} \right]^{-1} \left(\mathbf{C}_\mathbf{x}^{(\ell)} \right)^{-1}. \tag{60}$$

Substituting (60) into (35), we note that the LMSE estimator is given exactly by (17b), i.e., $\hat{\mathbf{x}}^{(\mathrm{LMSE})} = \boldsymbol{\mu}_\mathbf{x}$. Thus, the LMSE estimator is asymptotically optimal as $N \to \infty$.

5 Related Works

Fusion of estimates has been widely studied in many different areas. On the one hand, in wireless sensor networks the focus has been on distributed learning/estimation under communication constraints [20, 13] and the adaptation of methods developed for graphical models to distributed fusion [6]. Many different consensus, gossip or diffusion algorithms [12, 7, 5] have been developed, but they require a significant amount of communication that may constitute a burden in big data applications. On the other hand, a related field in the statistical literature is the combination of forecasts [17]. Indeed, the optimal linear combination for the single parameter case was already derived in [1, 2] and a Bayesian perspective was provided in [4]. However, there are two important differences

with respect to the scenario addressed here: (1) each forecaster is assumed to have access to the whole data set; (2) computational complexity is not considered an issue in those cases. Finally, there is currently a great interest in parallel Bayesian computation using Monte Carlo methods [19], and a few communication-free parallel Markov chain Monte Carlo (MCMC) algorithms have been developed [15, 11, 18]. However, none of them addresses the potentially large dimension of the optimal combiners.

6 Numerical Results

In this section we show the result of some preliminary experiments performed in [9]. We address the problem of positioning a target in the two-dimensional space of a wireless sensor network with only range measurements [10]. More specifically, we consider a random vector $\mathbf{X} = [X_1, X_2]^\top$ to denote the target's position in the $\mathbb{R}^2$ plane. The position of the target is then a specific realization $\mathbf{x}$. The measurements are obtained from 6 range sensors located at $\mathbf{h}_1 = [1, -8]^\top$, $\mathbf{h}_2 = [8, 10]^\top$, $\mathbf{h}_3 = [-15, -7]^\top$, $\mathbf{h}_4 = [-8, 1]^\top$, $\mathbf{h}_5 = [10, 0]^\top$ and $\mathbf{h}_6 = [0, 10]^\top$. The measurement equations are given by

$$Y_j = -20 \log \left(||\mathbf{x} - \mathbf{h}_j||^2 \right) + \Theta_j, \quad j = 1, \ldots, 6, \tag{61}$$

where $\Theta_j \sim \mathcal{N}(\theta_j | \mathbf{0}, \omega_j^2 \mathbf{I})$, with $\omega_j = 5$ for $j \in \{1, 2, 3\}$ and $\omega_j = 20$ for $j \in \{4, 5, 6\}$. We simulate $N = 6000$ observations from the model ($\frac{N}{6} = 1000$ observations from each of sensors) fixing $x_1 = 3.5$ and $x_2 = 3.5$. We consider a varying number of partial estimators L with $N_\ell = N/L$ for $1 \leq \ell \leq L$, and three scenarios for splitting the data:

Sc1 Exactly $\frac{N}{6L}$ measurements from each sensor are provided to each partial estimator.

Sc2 The first $L/2$ estimators contain an equal number of observations from the first 3 sensors (the best ones), whereas the remaining $L/2$ estimators work with measurements from the last 3 sensors (the noisiest ones).

Sc3 Measurements are randomly assigned to the estimators.

For each scenario, we run $M_C^{(\ell)} = 100$ MCMC independent parallel chains with length $T_C^{(\ell)} = 5000$, compute the MMSE estimates $\hat{x}_1^{(\ell)}$ and $\hat{x}_2^{(\ell)}$, and fuse these estimates into the final result. We compare the Equal Weights Fusion (EWF) method, where each estimator is given the same weight, $1/L$, and the three fusion methods described in Section 4. We repeat the experiments 50 times and average the results. The results, shown in Table 2 and Fig. 1, confirm the good performance of the SCME and ILMSE estimators, which outperform the naive EWF and show an MSE similar to the optimal and more costly LMSE. Note that the poor performance of all the estimators for small values of L is due to the slower convergence of the parallel chains when the number of data in the posterior is large (e.g., for $T_C^{(\ell)} = 20000$ the MSE decreases to 0.1624 when $L = 1$). This shows the importance of splitting the data even when a single estimator is able to deal with them. Finally, regarding the three scenarios considered, we note that the best performance is obtained in the second case (with $\mathrm{MSE}(\hat{\mathbf{x}}^{(\mathrm{LMSE})} | \vec{y}) = 0.0021$), i.e., splitting the data in separate filters according to their quality. This opens up the possibility of performing a "smart" division of the data in order to optimize the performance.

7 Conclusions and Future Lines

In this paper we have addressed the fusion of unbiased and uncorrelated partial minimum mean squared error (MMSE) estimators using two novel efficient linear combination schemes. The methods were tested through computer simulations by applying them to a localization problem with one target and six sensors whose measurements were processed using several parallel filters. The new fusion methods show a performance equivalent to the optimal linear combination with a reduced computational cost.

Experiment		N_ℓ								
Scenario	**Estimator**	6	12	30	60	240	600	1200	3000	6000
Sc1	EWF	0.0041	0.0049	0.0065	0.0090	0.0167	0.0590	0.1192	0.2899	0.5540
	SCMSE	0.0039	0.0046	0.0063	0.0089	0.0166	0.0587	0.1191	0.2899	
	ILMSE	0.0038	0.0046	0.0063	0.0089	0.0166	0.0586	0.1188	0.2886	
	LMSE	0.0037	0.0045	0.0062	0.0088	0.0165	0.0584	0.1183	0.2878	
Sc2	EWF	0.0087	0.0053	0.0064	0.0104	0.0343	0.0648	0.1681	0.3392	0.5540
	SCMSE	0.0057	0.0034	0.0047	0.0092	0.0328	0.0628	0.1623	0.3290	
	ILMSE	0.0052	0.0031	0.0043	0.0085	0.0304	0.0588	0.1521	0.3159	
	LMSE	0.0037	0.0021	0.0028	0.0057	0.0210	0.0410	0.1107	0.2406	
Sc3	EWF	0.0078	0.0061	0.0068	0.0092	0.0169	0.0587	0.1181	0.2877	0.5540
	SCMSE	0.0060	0.0053	0.0066	0.0091	0.0168	0.0584	0.1180	0.2877	
	ILMSE	0.0055	0.0051	0.0065	0.0090	0.0168	0.0583	0.1177	0.2867	
	LMSE	0.0051	0.0048	0.0064	0.0090	0.0167	0.0582	0.1174	0.2861	

Table 2: MSE (averaged over 50 independent runs) for the three scenarios and the four considered fusion methods.

.

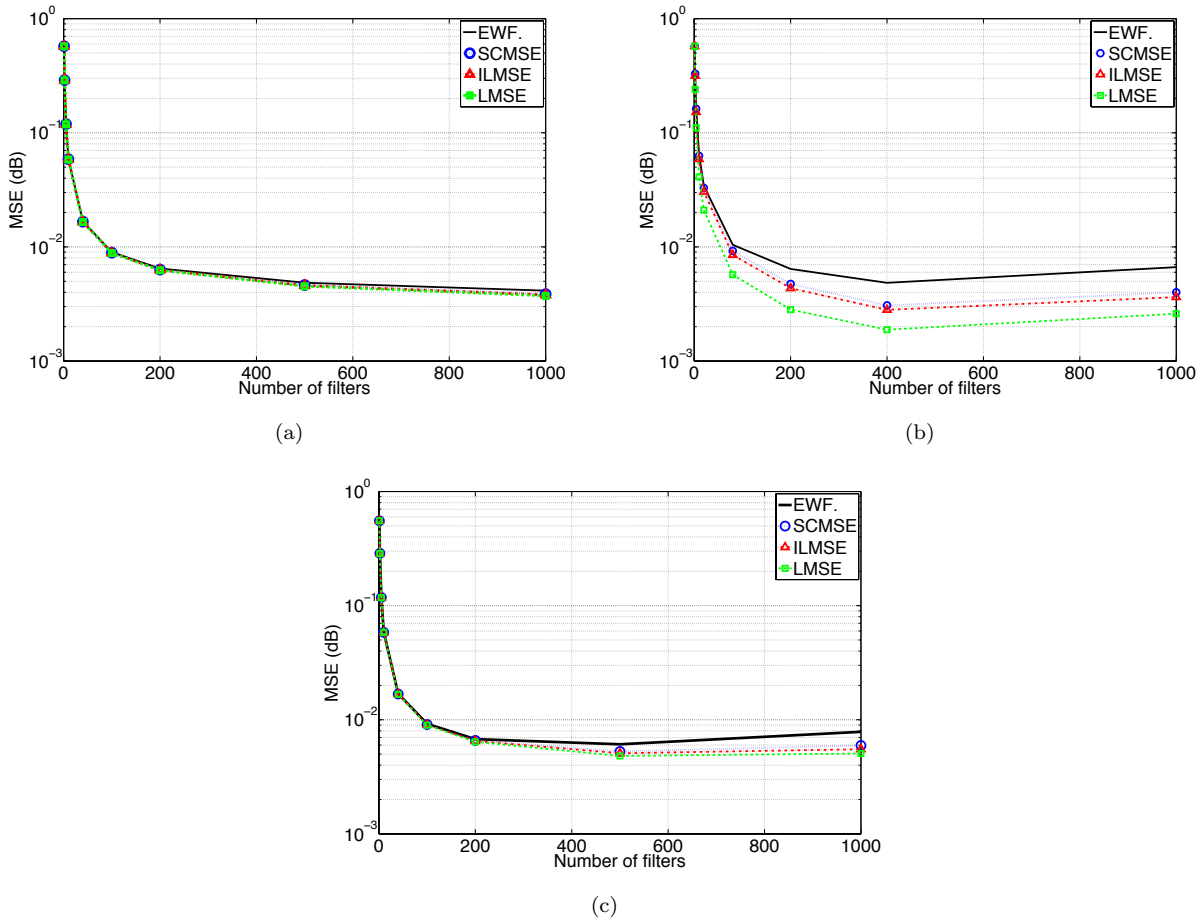

Figure 1: MSE as a function of L. **(a)** Scenario 1 (Sc1). **(b)** Scenario 2 (Sc2). **(c)** Scenario 3 (Sc3).

From a theoretical point of view, in future works we plan to address the fusion of biased and/or correlated

partial estimators,[10] particular applications that lead to simple fusion rules, non-linear fusion techniques and the development of fusion schemes where the partial Monte Carlo estimators are allowed to exchange a reduced amount of information. From a practical point of view, further simulations with larger amounts of data and higher-dimensional state spaces need to be performed.

A Estimation of a Scalar Parameter in AWGN

Let us consider the estimation of a constant parameter, $x \in \mathbb{R}$, contaminated by AWGN:

$$y[n] = x + w[n], \tag{62}$$

where $w[n] \sim \mathcal{N}(0, \sigma_w^2)$. Now, let us assume that we have N observations, s.t. $\mathbf{y} = [y[0], \ \ldots, \ y[N-1]]^\top$. The likelihood for this model is given by

$$\mathcal{L}(\mathbf{y}|x) = \frac{1}{(2\pi\sigma_w^2)^{N/2}} \exp\left(-\frac{1}{2\sigma_w^2}\sum_{n=0}^{N-1}(y[n]-x)^2\right). \tag{63}$$

This expression can be easily manipulated to express it as a Gaussian function of x:

$$\mathcal{L}(\mathbf{y}|x) = \frac{\exp(\Delta)}{(2\pi\sigma_w^2)^{N/2-1}} \times \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left(-\frac{1}{2\sigma_w^2}\left(x - \sum_{n=0}^{N-1} y[n]\right)^2\right), \tag{64}$$

where

$$\Delta = \sum_{n=0}^{N-1} y[n]^2 - \frac{1}{N}\left(\sum_{n=0}^{N-1} y[n]\right)^2. \tag{65}$$

If we consider a uniform prior for x, s.t. $p(x) \propto 1$,[11] then the MMSE estimator coincides with the maximum likelihood (ML) estimator:

$$\begin{aligned}
\hat{\mathbf{x}}^{(\mathrm{MMSE})} = \hat{\mathbf{x}}^{(\mathrm{ML})} = I(\mathbf{y}) &= \frac{1}{Z(\mathbf{y})}\int_{-\infty}^{\infty} x\ \mathcal{L}(\mathbf{y}|x)\ dx \\
&= \frac{1}{Z(\mathbf{y})} \times \frac{\exp(\Delta)}{(2\pi\sigma_w^2)^{N/2-1}} \times \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma_w^2}} \exp\left(-\frac{1}{2\sigma_w^2}\left(x - \sum_{n=0}^{N-1} y[n]\right)^2\right) \\
&\ldots = \frac{1}{Z(\mathbf{y})} \times \frac{\exp(\Delta)}{(2\pi\sigma_w^2)^{N/2-1}} \times \frac{1}{N}\sum_{n=0}^{N-1} y[n].
\end{aligned} \tag{66}$$

In order to obtain the final expression of $\hat{\mathbf{x}}^{(\mathrm{MMSE})}$, we need to obtain the partition function,

$$Z(\mathbf{y}) = \int_{-\infty}^{\infty} \mathcal{L}(\mathbf{y}|x)\ dx = \frac{\exp(\Delta)}{(2\pi\sigma_w^2)^{N/2-1}}. \tag{67}$$

Finally, inserting (67) into (66), the first part gets cancelled and we obtain the well-known MMSE/ML estimator for a constant scalar paramter in AWGN:

$$\hat{\mathbf{x}}^{(\mathrm{MMSE})} = \hat{\mathbf{x}}^{(\mathrm{ML})} = \frac{1}{N}\sum_{n=0}^{N-1} y[n]. \tag{68}$$

[10] The issue of combining biased and/or correlated partial estimators has an important practical interest. On the one hand, Monte Carlo estimators based on the importance sampling (IS) principle return asymptotically unbiased estimators, but the bias could be large for small values of ℓ. On the other hand, correlation among partial estimators may occur when the observations are not independent or when some data are assigned to multiple filters.

[11] Note that this is an improper prior, as $x \in \mathbb{R}$, and the normalizing constant can be included in the partition function.

Now, let us assume that we partition the whole dataset into L disjoint datasets containing the same number of observations $N_\ell = N/$ for $1 \leq \ell \leq L$, i.e.,

$$\mathbf{y}_\ell^\top = [y_\ell[0],\ \ldots,\ y_\ell[N/L-1]]^\top = [y[(\ell-1)N/L],\ \ldots,\ y[\ell N/L-1]]^\top,$$

and the full dataset can be constructed as the concatenation of the L partial datasets: $\mathbf{y} = [\mathbf{y}_1^\top,\ \ldots,\ \mathbf{y}_L^\top]^\top$. As a function of x, the ℓ-th partial likelihood can be expressed as

$$\mathcal{L}(\mathbf{y}_\ell|x) = \frac{\exp(\Delta_\ell)}{(2\pi\sigma_w^2)^{N/2L-1}} \times \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left(-\frac{1}{2\sigma_w^2}\left(x - \sum_{n=0}^{N/L-1} y_\ell[n]\right)^2\right), \tag{69}$$

where

$$\Delta_\ell = \sum_{n=0}^{N/L-1} y_\ell[n]^2 - \frac{1}{N/L}\left(\sum_{n=0}^{N/L-1} y_\ell[n]\right)^2. \tag{70}$$

Following the same reasoning as before, the ℓ-th partial MMSE/ML estimator is given by

$$\hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})} = \hat{\mathbf{x}}_\ell^{(\mathrm{ML})} = \frac{1}{Z_\ell(\mathbf{y}_\ell)} \times \frac{\exp(\Delta_\ell)}{(2\pi\sigma_w^2)^{N/2L-1}} \times \frac{1}{N/L} \sum_{n=0}^{N/L-1} y_\ell[n] = \frac{1}{N/L} \sum_{n=0}^{N/L-1} y_\ell[n], \tag{71}$$

with the last expression arising from the fact that

$$Z_\ell(\mathbf{y}_\ell) = \int_{-\infty}^{\infty} \mathcal{L}(\mathbf{y}_\ell|x)\ dx = \frac{\exp(\Delta_\ell)}{(2\pi\sigma_w^2)^{N/2L-1}}. \tag{72}$$

Hence, it is straightforward to see that

$$\hat{\mathbf{x}}^{(\mathrm{MMSE})} = \frac{1}{L}\sum_{\ell=1}^{L} \hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})}, \tag{73}$$

which corresponds to (12) with $\alpha_\ell = 1/L$ for $\ell = 1,\ \ldots,\ L$.

References

[1] J. M. Bates and C. W. Granger. The combination of forecasts. *Operational Research Quarterly*, 20(4):451–468, 1969.

[2] J. M. Bates and C. W. Granger. Some statistical results in the combination of forecasts. *Operational Research Quarterly*, 24:253–260, 1975.

[3] Dimitri P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1982.

[4] Robert F. Bordley. The combination of forecasts: A bayesian approach. *Journal of the Operational Research Society*, 33(2):171–174, 1982.

[5] Federico S. Cattivelli and Ali H. Sayed. Diffusion LMS strategies for distributed estimation. *IEEE Transactions on Signal Processing*, 58(3):1035–1048, 2010.

[6] Mujdat Cetin, Lei Chen, John W. Fisher III, Alexander T. Ihler, Randolph L. Moses, Martin J. Wainwright, and Alan S. Willsky. Distributed fusion in sensor networks. *IEEE Signal Processing Magazine*, 23(4):42–55, 2006.

[7] Alexandros G. Dimakis, Soummya Kar, José F. Moura, Michael G. Rabbat, and Anna Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010.

[8] Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York, NY (USA), 1986.

[9] David Luengo, Luca Martino, Víctor Elvira, and Mónica Bugallo. Efficient linear combination of partial Monte Carlo estimators. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 4–9 2015. Submitted.

[10] Luca Martino and Joaquín Míguez. Generalized rejection sampling schemes and applications in signal processing. *Signal Processing*, 90(11):2981–2995, November 2010.

[11] Willie Neiswanger, Chong Wang, and Eric Xing. Asymptotically exact, embarrassingly parallel MCMC. *arXiv:1311.4780v2*, pages 1–16, 21 Mar. 2014.

[12] Reza Olfati-Saber, J. Alex Fax, and Richard M. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.

[13] Joel B. Predd, Sanjeev R. Kulkarni, and H. Vincent Poor. Distributed learning in wireless sensor networks. *IEEE Signal Processing Magazine*, 23(4):56–69, 2006.

[14] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

[15] Steven L. Scott, Alexander W. Blocker, Fernando V. Bonassi, Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bayes and big data: The consensus Monte Carlo algorithm. In *EFaBBayes 250th conference*, volume 16, 2013.

[16] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[17] Kenneth F. Wallis. Combining forecasts – forty years later. *Applied Financial Economics*, 21(1–2):33–41, 2011.

[18] Xiangyu Wang and David B. Dunson. Parallelizing MCMC via Weierstrass sampler. *arXiv:1312.4605v2*, 25 May 2014.

[19] Darren J. Wilkinson. Parallel Bayesian computation. *Statistics Textbooks and Monographs*, 184, 2006.

[20] Jin-Jun Xiao, Alejandro Ribeiro, Zhi-Quan Luo, and Georgios B. Giannakis. Distributed compression-estimation using wireless sensor networks. *IEEE Signal Processing Magazine*, 23(4):27–41, 2006.