RESEARCH

Efficient Linear Fusion of Distributed MMSE Estimators for Big Data

David Luengo[1*], Luca Martino[2], Víctor Elvira[3] and Mónica Bugallo[4]

Abstract

Many signal processing applications require performing statistical inference on large datasets, where computational and/or memory restrictions become an issue. In this big data setting, computing an exact global centralized estimator is often unfeasible. Furthermore, even when approximate numerical solutions (e.g., based on Monte Carlo methods) working directly on the whole dataset can be computed, they may not provide a satisfactory performance either. Hence, several authors have recently started considering distributed inference approaches, where the data is divided among multiple workers (cores, machines or a combination of both). The computations are then performed in parallel and the resulting distributed or partial estimators are finally combined to approximate the intractable global estimator. In this paper, we focus on the scenario where no communication exists among the workers, deriving efficient linear fusion rules for the combination of the distributed estimators. Both a Bayesian perspective (based on the Bernstein-von Mises theorem and the asymptotic normality of the estimators) and a constrained optimization view are provided for the derivation of the linear fusion rules proposed. We concentrate on minimum mean squared error (MMSE) partial estimators, but the approach is more general and can be used to combine any kind of distributed estimators as long as they are unbiased. Numerical results show the good performance of the algorithms developed, both in simple problems where analytical expressions can be obtained for the distributed MMSE estimators, and in a wireless sensor network localization problem where Monte Carlo methods are used to approximate the partial estimators.

Keywords: big data; distributed estimation; minimum mean squared error (MMSE) estimators; Bayesian inference; Bernstein-von Mises theorem; Monte Carlo methods; linear fusion; constrained minimization

*Correspondence:
david.luengo@upm.es
[1]Dep. of Signal Theory and Communications, Technical Univ. of Madrid, Madrid, Spain
Full list of author information is available at the end of the article

1 Introduction

Estimation theory addresses the problem of inferring a set of unknown variables of interest given a collection of available data [1, 2]. This is a central problem in statistical signal processing, where a parametric model for the data is often assumed and its parameters have to be inferred from the observations [3, 4, 5]. Indeed, even non-parametric approaches typically have a reduced set of hyperparameters that have to be estimated from the data [6, 7, 8]. Unfortunately, determining the *global estimator* of these parameters using all the available information is often unfeasible or impractical for many real-world scenarios. Many current signal processing applications require performing statistical inference on large datasets, where the amount of data at hand imposes computational and/or storage constraints that impede the global estimation process [9]. Furthermore, even when approximate numerical solutions working directly on the whole dataset can be computed, they may not provide a satisfactory performance either. For example, Monte Carlo (MC) methods are often used to attain asymptotically exact estimators when closed-form analytical expressions cannot be otained [10, 11, 12]. However, large datasets pose a challenge for MC-based estimators, since the posterior density tends to concentrate

on a relatively small space as the number of data increases [13]. MC algorithms may have trouble locating this area (especially if the state space is also large) and thus can lead to a poor performance in practice.

An alternative to *global estimation* is dividing the available data into groups of manageable information, and distribute them among multiple workers (cores, machines or a combination of both). The computations are then performed in parallel (with or without communication among the different workers) and *distributed* or *partial* estimators of the unknown parameters are obtained. In this setting, two extreme situations may arise, namely the multi-core and the multi-machine scenarios [14]. On the one hand, in the *multi-core* case, the estimation is performed using several cores of a single machine (e.g., inside a graphics processing unit [GPU]) and communication among the cores can be considered costless [15, 16]. This approach allows for communication among workers, can provide significant speed-ups (if synchronization issues are properly addressed), and solves the computational cost problem, but not the memory/disk storage bottleneck. On the other hand, in the *multi-machine* case, the estimation is distributed among several machines (typically lying inside a large cluster), and the cost of inter-machine communications cannot be ignored. This approach can alleviate all the issues associated to big data signal processing (i.e., both computational and memory/storage issues), but requires each machine to work independently without any communication among workers (which typically communicate only to the central node at the beginning and the end of their tasks) [17]. Finally, note that a combination of both scenarios often occurs in practice (i.e., a large cluster where each machine may have several cores), thus resulting in situations where a moderate amount of communications may be acceptable.

In this paper, we focus on the scenario where no communication exists among the workers, deriving efficient linear fusion rules for the combination of the distributed minimum mean squared error (MMSE) estimators. The objective is thus finding an optimal combination of these distributed or partial estimators to achieve the performance of the global one. The fusion of different models or estimators has been widely studied in many different areas including control, signal processing, economics and communications. The literature on the subject is rather vast, and here we only mention the most important results related to the addressed problem.

On the one hand, a related field in the statistical literature is the combination of forecasts [18]. Indeed, the optimal linear combination for the single parameter case was already derived in [19, 20], a Bayesian perspective was provided in [21], and a general procedure to combine estimators in the multiple parameter case has been proposed very recently in [22]. However, there are two important differences with respect to the scenario addressed here: (1) each forecaster is assumed to have access to the whole dataset; (2) the computational complexity issue is not addressed. Therefore, problems related to the scarcity of data per estimator (when the number of data is large but the ratio data/workers is not so large), such as the so-called *small sample bias* [23], or the feasibility of the optimal combination rules when the number of parameters to be estimated is also large, have never been investigated in this context as far as we know.

On the other hand, in wireless sensor networks the focus has been on distributed learning/estimation under communication constraints [24, 25]. The optimal linear

fusion rule for the multi-dimensional case has also been derived in this context [25, 26], but the focus has been on developing optimal compression rules to restrict the amount of information being transmitted, rather than on obtaining efficient fusion schemes. However, this compression is not useful in the multi-machine learning scenario, since passing messages among multiple machines is expensive regardless of their size [14]. Distributed fusion approaches, obtained by adapting methods developed for graphical models, have also been proposed [27], as well as many different consensus, gossip or diffusion algorithms [28, 29, 30]. However, all of these methods require a significant amount of communication that constitutes a burden for multi-machine signal processing.

Finally, there is currently a great deal of interest in parallel Bayesian computation using MC methods [31], and a few communication-free parallel Markov chain Monte Carlo (MCMC) algorithms working on disjoint partial datasets have been developed following the so-called *embarrassingly parallel* architecture [32]. In [33], four alternatives were proposed to combine the samples drawn from the partial posteriors using either a Gaussian approximation or importance resampling. Then, [14] derived the optimal linear combination of weights required to obtain samples approximately from the full posterior, noting that the approach is optimal when both the full and the partial posteriors are Gaussian. This was followed by [34], where three different approaches to approximate the full posterior from the partial posteriors were proposed: a simple parametric approach, a non-parametric estimator and a semi-parametric method. At last, [35] proposed using the Weierstrass transform to improve the quality of the approximation to the full posterior. However, none of these previous works addresses the potentially large dimension of the optimal combiners. This issue has been initially tackled in [36]. In this paper we ellaborate on that work, providing a theoretical analysis of the proposed fusion rules, delving deeper into their underlying strengths and limitations, and performing more simulations to analyze their performance in practice.

1.1 Main Contributions

The main contribution of this work is the derivation of two novel efficient linear schemes for the fusion of the distributed or partial estimators. Although we focus on minimum mean squared error (MMSE) partial estimators throughout the paper, the proposed fusion schemes are independent from the specific approach followed to obtain those partial estimators (they are only assumed to be unbiased). The motivation comes from the optimal linear combination, which involves the calculation of one weighting matrix per partial estimator and thus may be too computationally demanding for large dimensional systems (both in number of unknowns and observations), as it requires as many weighting matrices (whose size depends quadratically on the number of unknowns) as partial estimators (whose number is typically a fraction of the number of observations). For instance, in a setting where the number of parameters to be estimated is D and the N observations available are equally distributed among L partial estimators, the optimal linear fusion approach requires computing one $D \times D$ matrix per partial estimator (L matrices and LD^2 parameters in total), which must be estimated from the partial dataset composed of N/L samples. In order to reduce the computational complexity, we propose two

linear approaches that require only a single weighting coefficient per partial estimator (i.e., L weights in total) and one weighting coefficient per parameter and partial estimator (i.e., LD weights in total), respectively.

Another important contribution of the paper is providing both a Bayesian perspective (based on the Bernstein-von Mises theorem and the asymptotic normality of the estimators) and a constrained optimization view for the derivation of all the linear fusion rules considered. These two complementary visions help to explain their good performance even when the normality assumption is not fulfilled. The optimal linear combination, derived first, provides the global MMSE estimator only when the partial MMSE estimators have a Gaussian distribution. Under certain regularity conditions, this is ensured by the Bernstein-von Mises theorem in the large-sample size limit for each partial estimator (i.e., when N/L is large). However, even when this theorem is not fulfilled and the partial estimators do not follow a Gaussian distribution, the optimal linear fusion rule provides the best linear unbiased estimator given the unbiased partial estimates. This explains the good performance of the optimal fusion rule observed in [14] for some cases where the underlying distributions were not Gaussian. The efficient linear fusion rules derived next can then be seen as the optimal restricted linear fusion rules corresponding to a single coefficient and a diagonal matrix, respectively.

Finally, we analyze the performance of all the fusion rules on several numerical examples. First, we perform a detailed study on simple examples, where exact closed-form expressions for the partial and the global estimators can be obtained. This allows us to rule out any approximation effects (e.g., due to slow convergence and poor mixing in MC methods) and analyze the effect of the number of samples, the number of estimators, the prior, and the dimensionality of the state space. Then, we apply the proposed algorithms to the problem of target localization in a wireless sensor network using measurements acquired by several sensors with different noise characteristics. In this scenario, MC partial estimators (based on parallel chains) are used to deal with the groups of measurements, showing that the performance of the novel fusion rules is close to that of the optimal fusion rule with only a fraction of its computational cost.

1.2 Organization

The reminder of the paper is structured as follows. The notation and the problem statement are provided first in Section 2. This is followed by Section 3, which briefly recalls the Bayesian framework to derive parameter estimators based on the Bayesian risk (Section 3.1), provides the optimal MMSE fusion rule for the Gaussian case (Section 3.2), discusses the asymptotic optimality of this fusion rule in other cases based on the asymptotic normality of the partial MMSE estimators as formulated by Bernstein-von Mises theorem (Section 3.3), and provides some hints on fusion rules for particular cases (Section 3.4). An alternative approach is then pursued in Section 4, where the optimal linear combination method is obtained by solving a constrained minimization problem (Section 4.1), and two novel efficient linear fusion rules are also derived following this approach (Sections 4.2 and 4.3). Several numerical experiments are analyzed and discussed in Section 5, first on a simple problem where analytical expressions for the partial MMSE estimators can

be obtained (Section 5.1) and then on a localization problem in wireless sensor networks, where MCMC methods have to be used to obtain the partial estimators (Section 5.2). Finally, some concluding remarks and future lines are provided in Section 6.

2 Problem Statement: Global vs. Partial Estimators

2.1 Exact Global Bayesian Estimators

Many applications in statistical signal processing require inferring a set of variables of interest or unknowns given a collection of observations or measurements. Let us consider a D-dimensional vector of unknowns, $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^D$, and let $\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^N$ be the collection of N i.i.d. observed data. From a Bayesian point of view, all the information about the unknown variables $\mathbf{x}$ is contained in the posterior probability density function (PDF), which is 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 $p(\mathbf{x}|\mathbf{y}) = \frac{1}{Z(\mathbf{y})}\pi(\mathbf{x}, \mathbf{y}) \propto \pi(\mathbf{x}, \mathbf{y})$.[1]

Let us assume a fixed model, where the likelihood and the priors are given and the posterior is thus automatically obtained by applying (1). The Bayesian inference problem is then solved by minimizing some risk function on the posterior PDF (see Section 3.1 for further details). For instance, it is well-known that the MMSE estimator corresponds to the conditional mean, i.e., the expected value of $\mathbf{x}$ w.r.t. the posterior PDF [1, 3, 4, 5],

$$\hat{\mathbf{x}}^{(\mathrm{MMSE})} = \mathbb{E}(\mathbf{x}|\mathbf{y}) = \int_{\mathcal{X}} \mathbf{x}\, p(\mathbf{x}|\mathbf{y}) d\mathbf{x}, \tag{3}$$

whereas the maximum a posteriori (MAP) estimator corresponds to the location of the highest mode in the posterior PDF,

$$\hat{\mathbf{x}}^{(\mathrm{MAP})} = \arg\max_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}|\mathbf{y}). \tag{4}$$

2.2 Asymptotically Exact Global Estimators: Monte Carlo-Based Approaches

Unfortunately, the direct computation of either (3) or (4) exactly is unfeasible in most problems of interest, especially for high-dimensional scenarios (i.e., for large

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

values of D). In those cases, a practical solution consists of using an MC approach to compute an asymptotically exact approximation of the desired estimator. MC-based algorithms are designed to provide an efficient approximation to some moment of $\mathbf{x}$ (i.e., an integral measure w.r.t. the target PDF),

$$I_{f(\mathbf{x})}(\mathbf{y}) = \int_{\mathcal{X}} f(\mathbf{x}) p(\mathbf{x}|\mathbf{y}) d\mathbf{x} = \frac{1}{Z(\mathbf{y})} \int_{\mathcal{X}} f(\mathbf{x}) \pi(\mathbf{x}, \mathbf{y}) d\mathbf{x}, \tag{5}$$

where $f(\mathbf{x})$ can be any integrable function of $\mathbf{x}$, and the unknown partition function is given by

$$Z(\mathbf{y}) = I_1(\mathbf{y}) = \int_{\mathcal{X}} \pi(\mathbf{x}, \mathbf{y}) d\mathbf{x}. \tag{6}$$

Monte Carlo approaches can be divided in two large families of methods: Markov chain Monte Carlo (MCMC) and importance sampling (IS). On the one hand, MCMC algorithms are based on sampling from a Markov chain whose stationary density is the target PDF, $\pi(\mathbf{x})$. Candidate samples are drawn from a proposal distribution $q(\mathbf{x})$, and they are either accepted or rejected according to some proper rule. After an initial "burn-in" period, it can be assumed that the chain has converged and the accepted samples are distributed according to the target, $\pi(\mathbf{x})$. Let us assume that we have M random samples drawn from the target PDF, $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}$ with $\mathbf{x}^{(m)} \sim \pi(\mathbf{x})$ for $m = 1, \ldots, M$. Then, MCMC-based approaches construct a numerical approximation to (5) and (6) as a sum of the function $f(\mathbf{x})$ evaluated at those samples,

$$\hat{I}^{(\mathrm{MCMC})}_{f(\mathbf{x})}(\mathbf{y}) = \frac{1}{M} \sum_{m=1}^{M} f(\mathbf{x}^{(m)}), \tag{7}$$

On the other hand, importance sampling (IS) approaches accept all the samples drawn from $q(\mathbf{x})$, weighting them appropriately according to their "quality". Thus, the numerical approximation to (5) and (6) corresponds now to a weighted sum of these samples,

$$\hat{I}^{(\mathrm{IS})}_{f(\mathbf{x})}(\mathbf{y}) = \sum_{m=1}^{M} w_m f(\mathbf{x}^{(m)}), \tag{8}$$

$$\hat{I}^{(\mathrm{IS})}_{1}(\mathbf{y}) = \sum_{m=1}^{M} w_m. \tag{9}$$

with the weights w_m depending on the specific approach followed. In the classical IS approach, the weights are given by the ratio between the target and the proposal evaluated at each sample (i.e., $w_m = \frac{\pi(\mathbf{x}^{(m)})}{q(\mathbf{x}^{(m)})}$), but other approaches to calculate the weights (such as the deterministic mixture weighting scheme) are possible [37].

2.3 Distributed Partial Bayesian Estimators

In big data problems, we cannot deal with the whole data set globally due to computational and/or memory restrictions.[2] A natural solution is splitting the data into L disjoint groups/clusters, so that the ℓ-th cluster ($1 \leq \ell \leq L$) only has access to N_ℓ samples. Then, we can obtain the partial MMSE estimator for each cluster (i.e., the MMSE estimator of $\mathbf{x}$ given all the data available to the ℓ-th estimator, $\mathbf{y}_\ell$) as

$$\hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})} = \mathbb{E}(\mathbf{x}|\mathbf{y}_\ell) = \int_{\mathcal{X}} \mathbf{x}\; p_\ell(\mathbf{x}|\mathbf{y}_\ell) d\mathbf{x}, \tag{10}$$

where $p_\ell(\mathbf{x}|\mathbf{y}_\ell) = \frac{1}{Z_\ell(\mathbf{y}_\ell)}\pi_\ell(\mathbf{x}, \mathbf{y}_\ell)$ is the partial posterior associated to the ℓ-th dataset (see Table 1 for a summary of the notation used throughout the paper). The goal is obtaining the global MMSE estimator, $\hat{\mathbf{x}}^{(\mathrm{MMSE})}$, from the set of partial MMSE estimators, $\{\hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})}\}_{\ell=1}^{L}$.[3]

In this paper we consider only the communication-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 in general the exact global MMSE estimator is a non-linear function of the whole dataset and cannot be attained from the partial MMSE estimators. A particular case where the exact global MMSE estimator can be obtained from the partial MMSE estimators occurs when both the global and the partial posteriors have Gaussian PDFs. In this case, it can be shown (see Section 3.2) that the global MMSE estimator is a weighted linear combination of the partial MMSE estimators:

$$\hat{\mathbf{x}}^{(\mathrm{MMSE})} = \sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell \hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})}, \tag{11}$$

where $\mathbf{\Lambda}_\ell$ is a $D \times D$ weighting matrix. When the conditions for the Bernstein-von Mises theorem are fulfilled, all the posterior PDFs are Gaussian and (11) becomes asymptotically optimal, as discussed in Section 3.3. However, even when the Bernstein-von Mises theorem does not hold, the linear combination of Eq. (11) can be a good fusion rule, since it corresponds to the best linear unbiased estimator of $\hat{\mathbf{x}}^{(\mathrm{MMSE})}$ given $\hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})}$ for $\ell = 1, \ldots, L$, as shown in Section 4.1. In the following sections we discuss all these issues and provide more efficient linear fusion rules (Sections 4.2 and 4.3), which correspond to restricted versions of the best linear unbiased estimator and may be optimal under certain circumstances.

[2] Even when we can deal with the whole data set globally, splitting it into L data sets may be more efficient and lead to a better performance. This is due to the fact that the posterior PDF tends to become more "peaky" as the number of data increases, thus rendering the inference process harder, especially for high-dimensional scenarios.

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

Table 1 Summary of the Notation.

$\mathbf{x}$	Unknown parameters to be estimated.
D	Number of unknowns (i.e., dimension of $\mathbf{x}$).
$\mathbf{y}$	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.
$\mathbf{y}_\ell$	Data set for the ℓ-th estimator.
$p(\mathbf{x}\|\mathbf{y})$	Global posterior PDF.
$p_\ell(\mathbf{x}\|\mathbf{y}_\ell)$	Partial posterior PDF for the ℓ-th estimator.
$\pi(\mathbf{x},\mathbf{y})$	Global target PDF.
$\pi_\ell(\mathbf{x},\mathbf{y}_\ell)$	Partial target PDF for the ℓ-th estimator.
$Z(\mathbf{y})$	Global partition function.
$Z_\ell(\mathbf{y}_\ell)$	Partial partition function for the ℓ-th estimator.

3 Optimal Linear Fusion: A Bayesian Perspective

3.1 Bayesian Risk

From a Bayesian point of view, the problem of finding an optimal estimator can be formulated as the minimization of a given risk function. Let us define the *Bayesian risk* as

$$R(\hat{\mathbf{x}}) = \int_{\mathcal{Y}}\int_{\mathcal{X}} C(\mathbf{x},\hat{\mathbf{x}})p(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y} = \int_{\mathcal{Y}} r(\hat{\mathbf{x}})p(\mathbf{y})d\mathbf{y}, \tag{12}$$

where $\hat{\mathbf{x}}$ can be any estimator of $\mathbf{x}$,

$$r(\hat{\mathbf{x}}) = \int_{\mathcal{X}} C(\mathbf{x},\hat{\mathbf{x}})p(\mathbf{x}|\mathbf{y})d\mathbf{x}, \tag{13}$$

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 (12) or (13) is equivalent. Now, let us consider the quadratic cost,

$$C(\mathbf{x},\hat{\mathbf{x}}) = (\hat{\mathbf{x}}-\mathbf{x})^\top(\hat{\mathbf{x}}-\mathbf{x}), \tag{14}$$

which is the most common cost function for regression problems. Then, (13) becomes

$$r(\hat{\mathbf{x}}) = \mathrm{MSE}(\hat{\mathbf{x}}|\mathbf{y}) = \int_{\mathcal{X}} (\hat{\mathbf{x}}-\mathbf{x})^\top(\hat{\mathbf{x}}-\mathbf{x})p(\mathbf{x}|\mathbf{y})d\mathbf{x}, \tag{15}$$

and the optimal estimator corresponds to the MMSE estimator, which is given by Eq. (3):

$$\hat{\mathbf{x}}^{(\mathrm{MMSE})} = \mathbb{E}(\mathbf{x}|\mathbf{y}) = \int_{\mathcal{X}} \mathbf{x}\; p(\mathbf{x}|\mathbf{y})d\mathbf{x}.$$

Let us remark that, in the Bayesian literature, $r(\hat{\mathbf{x}})$, as given by (15), is usually known as the *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 (14) in (12):

$$\mathrm{MSE}(\hat{\mathbf{x}}) = \int_{\mathcal{Y}} \int_{\mathcal{X}} (\hat{\mathbf{x}} - \mathbf{x})^\top (\hat{\mathbf{x}} - \mathbf{x}) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}. \tag{16}$$

Hence, strictly speaking Eq. (15) does not correspond to the Bayesian MSE. However, by assuming that the data are fixed, we can remove the outer integral in (16) 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 refer to it just as the MSE. Thus, whenever we mention the MSE in the sequel we refer to the conditional MSE as defined by (15).

3.2 Gaussian Estimators: Optimal Fusion Rule

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 have Gaussian densities with means equal to the true parameter vector $\mathbf{x}$ and covariance matrices $\mathbf{C}_{\mathbf{x}}^{(\ell)}$. Then, the full posterior is

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &= \prod_{\ell=1}^{L} \mathcal{N}(\hat{\mathbf{x}}_\ell | \mathbf{x}, \mathbf{C}_{\mathbf{x}}^{(\ell)}) \\ &= \prod_{\ell=1}^{L} (2\pi)^{-D/2} |\mathbf{C}_{\mathbf{x}}^{(\ell)}|^{-1/2} \\ &\times \exp\left(-\frac{1}{2} \sum_{\ell=1}^{L} (\hat{\mathbf{x}}_\ell - \mathbf{x})^\top \left(\mathbf{C}_{\mathbf{x}}^{(\ell)} \right)^{-1} (\hat{\mathbf{x}}_\ell - \mathbf{x}) \right). \end{aligned} \tag{17}$$

It is straightforward to see that

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) =& (2\pi)^{-D/2} |\mathbf{C}_{\mathbf{x}}|^{-1/2} \\ & \times \exp\left(-\frac{1}{2} (\hat{\mathbf{x}} - \boldsymbol{\mu}_{\mathbf{x}})^\top \mathbf{C}_{\mathbf{x}}^{-1} (\hat{\mathbf{x}} - \boldsymbol{\mu}_{\mathbf{x}}) \right), \end{aligned} \tag{18}$$

where

$$\mathbf{C}_{\mathbf{x}} = \left[\sum_{\ell=1}^{L} \left(\mathbf{C}_{\mathbf{x}}^{(\ell)} \right)^{-1} \right]^{-1}, \tag{19a}$$

$$\boldsymbol{\mu}_{\mathbf{x}} = \mathbf{C}_{\mathbf{x}} \sum_{\ell=1}^{L} \left(\mathbf{C}_{\mathbf{x}}^{(\ell)} \right)^{-1} \hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})}. \tag{19b}$$

Hence, the global MMSE estimator, which corresponds to the mean of the full posterior, is finally given by

$$\hat{\mathbf{x}}^{(\mathrm{MMSE})} = \sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell \hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})}, \tag{20}$$

with

$$\mathbf{\Lambda}_\ell = \mathbf{C}_\mathbf{x} \left(\mathbf{C}_\mathbf{x}^{(\ell)}\right)^{-1} = \left[\sum_{k=1}^{L} \left(\mathbf{C}_\mathbf{x}^{(k)}\right)^{-1}\right]^{-1} \left(\mathbf{C}_\mathbf{x}^{(\ell)}\right)^{-1}. \tag{21}$$

3.3 Asymptotic Normality: Bernstein-von Mises Theorem

The Bernstein-von Mises (a.k.a. Bayesian central limit) theorem states that, under suitable regularity conditions, a posterior PDF converges to a Gaussian PDF as the number of samples tends to infinity [38, 39]. Applying this result to the partial posterior PDFs, we have

$$p_\ell(\mathbf{x}|\mathbf{y}_\ell) \to \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x}^{(\ell)}, \mathbf{C}_\mathbf{x}^{(\ell)}) \qquad \text{as} \quad N_\ell \to \infty, \tag{22}$$

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

$$\begin{aligned} \mathbf{C}_\mathbf{x}^{(\ell)} &= \mathbb{E}\left((\hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})} - \mathbf{x})(\hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})} - \mathbf{x})^\top\right) \\ &= \int_\mathcal{X} (\hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})} - \mathbf{x})(\hat{\mathbf{x}}_\ell^{(\mathrm{MMSE})} - \mathbf{x})^\top p_\ell(\mathbf{x}|\mathbf{y}_\ell) d\mathbf{x}. \end{aligned} \tag{23}$$

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.,

$$p(\mathbf{x}|\mathbf{y}) = \prod_{\ell=1}^{L} p_\ell(\mathbf{x}|\mathbf{y}_\ell) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x}, \mathbf{C}_\mathbf{x}) \qquad \text{as} \quad N \to \infty, \tag{24}$$

with $\mathbf{C}_\mathbf{x}$ and $\mu_\mathbf{x}$ given by (19a) and (19b), respectively. In the context of distributed MC algorithms, Eq. (20) has been already proposed in [14] to combine samples from the partial posteriors in order to obtain approximate samples from the full posterior. This approach has also been exploited in [34] to obtain asymptotically exact samples from the global posterior by sampling from a multivariate Gaussian whose covariance matrix and mean vector are given by (19a) and (19b), respectively.

3.4 Particular Cases

Note that Eq. (20) requires computing a $D \times D$ weight matrix, given by (21), for each of the L estimators. This implies computing up to $D^2 L$ weights, which may

be unfeasible (or at least very costly from a computational/storage point of view) when D and/or L is large. 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

$$\mathbf{C}_{\mathbf{x}}^{(\ell)} = \mathrm{diag}\left(\sigma_{\ell,1}^2,\ \ldots,\ \sigma_{\ell,D}^2\right), \tag{25}$$

with

$$\sigma_{\ell,d}^2 = \int_{\mathcal{X}_d} (\hat{x}_{\ell,d} - x_d)^2 p(x_d|\mathbf{y}_\ell) dx_d \tag{26}$$

for $d = 1, \ldots, D$. In this scenario, the optimal weight matrix becomes

$$\mathbf{\Lambda}_\ell = \mathrm{diag}\left(\alpha_{\ell,1}^2,\ \ldots,\ \alpha_{\ell,D}^2\right), \tag{27}$$

with

$$\alpha_{\ell,d} = \frac{\sigma_{\ell,d}^{-2}}{\sum_{k=1}^{L} \sigma_{k,d}^{-2}}. \tag{28}$$

Note that only D parameters are required for each of the L estimators in this case (i.e., DL parameters in total). 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 in total), which can be obtained by averaging (28) over the set of all the parameters:

$$\alpha_\ell = \frac{1}{D}\sum_{d=1}^{D} \alpha_{\ell,d} = \frac{1}{D}\sum_{d=1}^{D} \left(\sum_{k=1}^{L} \sigma_{k,d}^{-2}\right)^{-1} \sigma_{\ell,d}^{-2}. \tag{29}$$

This corresponds to the best isotropic Gaussian approximation of the full posterior. Furthermore, when the partial estimators have the same variance for all the parameters (i.e., $\sigma_{\ell,d}^{-2} = \sigma_\ell^{-2}$ for $1 \leq d \leq D$), then (29) becomes

$$\alpha_\ell = \frac{\sigma_\ell^{-2}}{\sum_{k=1}^{L} \sigma_k^{-2}}, \tag{30}$$

which corresponds to the optimal weights. Finally, when all the covariance matrices of the partial estimators are equal, we simply have $\alpha_\ell = 1/L$.

4 Alternative Approach: Constrained Minimization

4.1 General Case: Optimal Linear Combination

Let us consider the most general linear combination of estimators,

$$\hat{\mathbf{x}} = \sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell \hat{\mathbf{x}}_\ell, \tag{31}$$

where $\hat{\mathbf{x}}_\ell$ can be any partial estimator (not necessarily the MMSE estimator) based on the ℓ-th partial dataset, $\mathbf{y}_\ell$, and $\hat{\mathbf{x}}$ is the corresponding global estimator obtained by linearly combining all those partial estimators. In this case, assuming that all the partial estimators are unbiased, the mean of the global estimator is given by

$$\mathbb{E}(\hat{\mathbf{x}}) = \sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell \mathbb{E}(\hat{\mathbf{x}}_\ell) = \left(\sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell \right) \mathbf{x}. \tag{32}$$

Thus, in order to obtain an unbiased global estimator we need to impose the following condition:

$$\sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell = \mathbf{I}. \tag{33}$$

The covariance matrix of the global estimator can be expressed as a function of the partial estimators, $\mathbf{C}_\mathbf{x}^{(\ell)}$, and the weight matrices, $\mathbf{\Lambda}_\ell$, as

$$\mathbf{C}_\mathbf{x} = \sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell \mathbf{C}_\mathbf{x}^{(\ell)} \mathbf{\Lambda}_\ell^\top. \tag{34}$$

The MSE of the global estimator is then given by

$$\mathrm{MSE}(\hat{\mathbf{x}}|\mathbf{y}) = \mathrm{Tr}\left(\mathbf{C}_\mathbf{x}\right) = \sum_{\ell=1}^{L} \mathrm{Tr}\left(\mathbf{\Lambda}_\ell \mathbf{C}_\mathbf{x}^{(\ell)} \mathbf{\Lambda}_\ell^\top \right), \tag{35}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix.

Now, in order to obtain the unbiased global estimator that minimizes the MSE, we need to solve the following constrained optimization problem:

$$\mathbf{\Lambda}^* = \arg\min_{\mathbf{\Lambda}} \sum_{\ell=1}^{L} \mathrm{Tr}\left(\mathbf{\Lambda}_\ell \mathbf{C}_\mathbf{x}^{(\ell)} \mathbf{\Lambda}_\ell^\top \right), \tag{36a}$$

$$\text{s.t.} \qquad \sum_{\ell=1}^{L} \mathbf{\Lambda}_\ell = \mathbf{I}, \tag{36b}$$

where $\mathbf{\Lambda} = [\mathbf{\Lambda}_1, \ \ldots, \ \mathbf{\Lambda}_L]^\top$. Since (36a) and (36b) correspond to a convex optimization problem, by applying the method of the Lagrange multipliers, it can be shown that the solution for each of the weighting matrices is simply given by Eq. (21):

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

Substituting this expression in (31), we note that the optimal linear MMSE (LMSE) fusion rule is given exactly by (20), i.e., $\hat{\mathbf{x}}^{(\mathrm{LMSE})} = \boldsymbol{\mu}_\mathbf{x}$, regardless of the approach followed to derive the partial estimators.

4.2 Particular Case: Single Coefficient

Let us consider the particular case in which a single coefficient per estimator is used to construct the global estimator:

$$\hat{\mathbf{x}} = \sum_{\ell=1}^{L} \alpha_\ell \hat{\mathbf{x}}_\ell, \tag{37}$$

which is obtained by setting $\mathbf{\Lambda}_\ell = \alpha_\ell \mathbf{I}$ in (31). Clearly this will provide a suboptimal solution in general, but it is a fast and low-cost solution for the combination of estimators, and we can easily 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 (37) is

$$\mathbb{E}(\hat{\mathbf{x}}) = \sum_{\ell=1}^{L} \alpha_\ell \mathbb{E}(\hat{\mathbf{x}}_\ell) = \left(\sum_{\ell=1}^{L} \alpha_\ell \right) \mathbf{x}. \tag{38}$$

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

$$\sum_{\ell=1}^{L} \alpha_\ell = 1. \tag{39}$$

On the other hand, the covariance matrix for the global estimator is given by

$$\mathbf{C}_\mathbf{x} = \sum_{\ell=1}^{L} \alpha_\ell^2 \mathbf{C}_\mathbf{x}^{(\ell)}, \tag{40}$$

and the MSE can be expressed as

$$\mathrm{MSE}(\hat{\mathbf{x}}|\mathbf{y}) = \mathrm{Tr}(\mathbf{C}_\mathbf{x}) = \sum_{\ell=1}^{L} \alpha_\ell^2 \mathrm{Tr}\left(\mathbf{C}_\mathbf{x}^{(\ell)} \right), \tag{41}$$

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

$$\mathrm{Tr}(\mathbf{C}_\mathbf{x}) = \sum_{d=1}^{D} \mathbf{C}_\mathbf{x}[d,d] = \sum_{d=1}^{D} \sigma_{x_d}^2, \tag{42}$$

with $\sigma_{x_d}^2 = \mathbb{E}((\hat{x}_d - x_d)^2)$, and $\mathrm{Tr}(\mathbf{C}_\mathbf{x}^{(\ell)})$ denotes the trace of the ℓ-th partial covariance matrix:

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

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

The goal is finding the set of α_ℓ that minimizes (41), subject to Eq. (39) in order to obtain an unbiased estimator. Hence, 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) correspond again to a convex optimization problem. Thus, by applying once more the method of the Lagrange multipliers, it can be shown that the single coefficient MMSE (SCMSE) fusion rule is given by

$$\begin{aligned}\hat{\mathbf{x}}^{(\mathrm{SCMSE})} &= \sum_{\ell=1}^{L} \frac{T_\ell^{-1}}{\sum_{k=1}^{L} T_k^{-1}} \hat{\mathbf{x}}_\ell \\ &= \sum_{\ell=1}^{L} \frac{\left[\mathrm{MSE}(\hat{\mathbf{x}}_\ell|\mathbf{y}_\ell)\right]^{-1}}{\sum_{k=1}^{L} \left[\mathrm{MSE}(\hat{\mathbf{x}}_k|\mathbf{y}_k)\right]^{-1}} \hat{\mathbf{x}}_\ell.\end{aligned} \tag{45}$$

4.3 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 (45) involve the trace of the partial covariance matrices, we introduce an independent linear minimum mean squared estimator (ILMSE) where $\boldsymbol{\Lambda}_\ell = \mathrm{diag}(\alpha_{\ell,1},\ \ldots,\ \alpha_{\ell,D})$. This approach leads to an independent estimation of each of the D unknowns:

$$\hat{x}_d^{(\mathrm{ILMSE})} = \sum_{\ell=1}^{L} \alpha_{\ell,d}\ \hat{x}_{\ell,d}^{(\mathrm{MMSE})}, \tag{46}$$

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 (46) can be obtained by solving D single parameter constrained optimization problems:

$$\boldsymbol{\alpha}_d^* = \arg\min_{\boldsymbol{\alpha}_d} \sum_{\ell=1}^{L} \alpha_{\ell,d}^2 C_{x_d}^{(\ell)}, \tag{47a}$$

$$\text{s.t.} \qquad \sum_{\ell=1}^{L} \alpha_{\ell,d} = 1, \tag{47b}$$

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

$$\alpha_{\ell,d} = \frac{\left[\mathrm{MSE}(\hat{x}_{\ell,d}^{(\mathrm{MMSE})}|\mathbf{y}_\ell)\right]^{-1}}{\sum_{k=1}^{L} \left[\mathrm{MSE}(\hat{x}_{k,d}^{(\mathrm{MMSE})}|\mathbf{y}_k)\right]^{-1}}. \tag{48}$$

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.

5 Numerical Experiments

5.1 Simple Example: Gamma Distributions

Let us consider first a simple one-dimensional example, where exact calculations may be performed. This allows us to rule out any potential issue with the underlying MC methods typically used to approximate the MMSE estimators (e.g., slow convergence and poor mixing), and concentrate on the performance of the proposed fusion rules. Let us assume that we have N i.i.d. observations distributed according to a Gamma PDF with known shape parameter $\alpha > 0$ and unknown rate parameter $\beta > 0$. Then, the likelihood is given by

$$\mathcal{L}(\mathbf{y}|\beta) = \prod_{n=1}^{N} \mathcal{L}(y_n|\beta), \tag{49}$$

with

$$\mathcal{L}(y_n|\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y_n^{\alpha-1} \exp(-\beta y_n). \tag{50}$$

The conjugate prior is also a Gamma PDF over β with shape parameter $\alpha_0 > 0$ and rate parameter $\beta_0 > 0$, and thus the global posterior density is another Gamma PDF with parameters $\alpha^* = \alpha_0 + \alpha$ and $\beta^* = \beta_0 + \frac{1}{N}\sum_{n=1}^{N} y_n$. Hence, the global MMSE estimator is

$$\hat{\beta}^{(\mathrm{MMSE})} = \frac{\alpha^*}{\beta^*} = \frac{\alpha_0 + \alpha}{\beta_0 + \frac{1}{N}\sum_{n=1}^{N} y_n}, \tag{51}$$

and its variance is given by

$$\sigma^2_{\beta} = \frac{\alpha^*}{(\beta^*)^2} = \frac{\alpha_0 + \alpha}{\left(\beta_0 + \frac{1}{N}\sum_{n=1}^{N} y_n\right)^2}. \tag{52}$$

For the distributed estimators, the partial MMSE estimates and their variances are still given respectively by (51) and (52), but taking the sum only over the N_ℓ samples available to each of the ℓ estimators.

We are interested now in analyzing the effect of the sample size, the number of partial estimators and the number of samples per estimator. Therefore, we test $N \in \{10^3, 10^4, 10^5, 10^6, 5 \cdot 10^6\}$ with an equal number of samples per partial estimator ranging from $N_\ell = 1$ (i.e., as many partial estimators as observations) up to $N_\ell = N$ (i.e., a single estimator that corresponds to the global estimator). For each case 1000 simulations are performed (except for $N = 10^6$ and $N = 5 \cdot 10^6$, where only 100 simulations have been performed) to average the results.

Figure 1 shows the typical performance of the optimal linear fusion rule (since we only have one parameter, all the fusion rules discussed in the paper are equivalent),

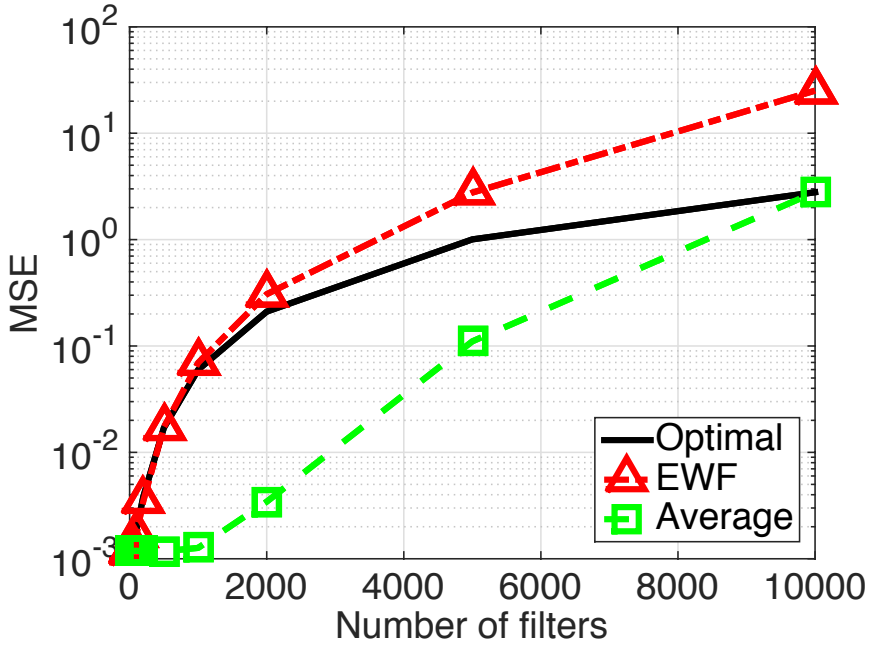

Figure 1 Simple Example. Conditional MSE as a function of L for the Gamma posterior PDF using an improper prior.

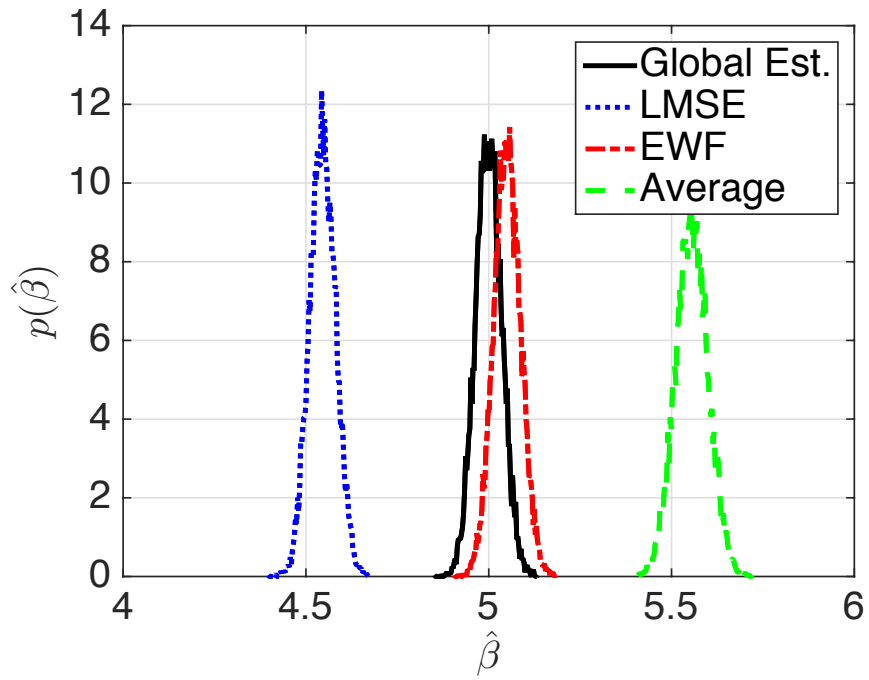

Figure 2 Simple Example. Posterior density of the estimator for the three fusion rules considered compared to the posterior of the global estimator.

the equal weights fusion (EWF) rule (that assigns the same weight to all the partial estimators) and an empirical estimator that combines the optimal and the EWF

estimates at the fusion center. In this example, the true parameters are $\alpha = 2$ and $\beta = 5$, and an improper prior is used by setting $\alpha_0 = \beta_0 = \epsilon$ with $\epsilon \to 0$. First of all, note that the optimal linear fusion rule performs better than the EWF (as expected), especially when the number of partial estimators (a.k.a. filters) increases. The unexpected result is that combining the optimal fusion strategy and the EWF approach leads to a better performance than any of the two individual strategies. The reason for this good performance can be seen in Figure 2, which shows the estimated posterior PDFs of the estimators for the three fusion rules considered compared to the posterior of the global estimator. It can be seen that the optimal linear fusion rule introduces a negative bias, whereas the EWF introduces a positive bias with approximately the same magnitude. Therefore, combining the two estimator leads to an average estimator with a reduced bias and thus a better performance.

The second important issue in Figure 1 is related to the increase of the MSE as the number of partial estimators increases. This is precisely due to the fact that the bias increases as the number of samples per partial estimator decreases (i.e., as the number of partial estimator increases for a fixed number of data). This bias is caused by the mismatch between the "true" prior (in this case a delta centered at the true value $\beta = 5$) and the prior assumed by the model. In order to see this, Figure 3 shows the evolution of the MSE with the number of filters using a narrow prior (obtained setting $\beta_0 = \frac{\beta}{\varepsilon}$ and $\alpha_0 = \beta \times \beta_0$ for $\varepsilon = 0.01$) centered around the true value of β. In this case all the estimators are unbiased and the MSE decreases as we increase the number of partial estimators. These results, in an example where the exact MMSE estimator can be obtained, highlights the importance of the prior in the Bayesian distributed inference approach.

Finally, Table 5.1 provides the complete picture regarding the evolution of the MSE with the number of data and the number of data per partial estimator for all the fusion rules considered. On the one hand, note that a minimum amount of samples per estimator are required in order to attain a performance that decreases as a function of N for the optimal linear fusion and the EWF. Otherwise, the bias dominates and nothing is gained by increasing N. On the other hand, note the excellent behaviour of the average fusion rule for all the cases.

5.2 Localization in a Wireless Sensor Network

In this section, we address the problem of positioning a static target in the two-dimensional space of a wireless sensor network using only range measurements. 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

$$Y_j = -20 \log\left(||\mathbf{x} - \mathbf{h}_j||^2\right) + \Theta_j, \quad j = 1, \ldots, 6, \tag{53}$$

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

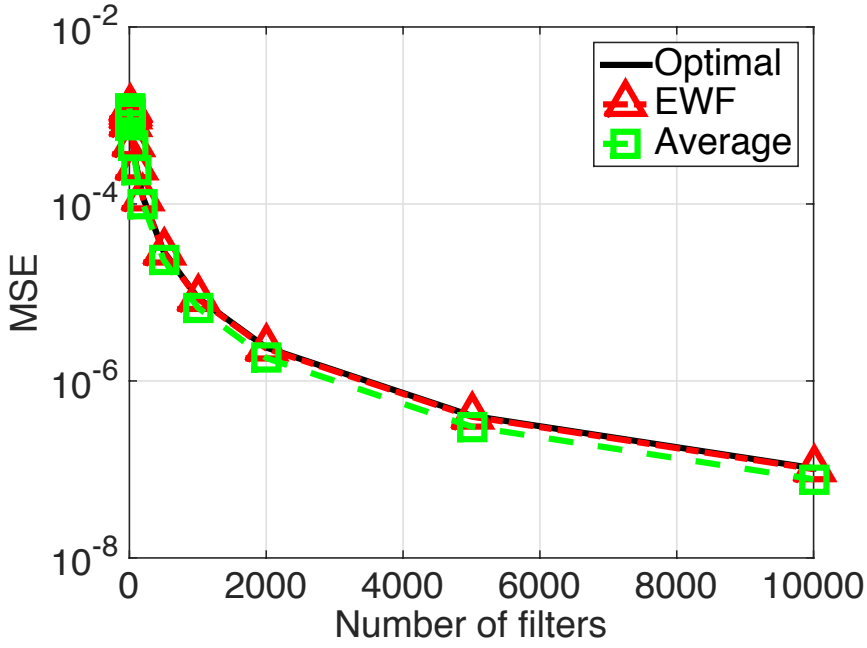

Figure 3 Simple Example. Conditional MSE as a function of L for the Gamma posterior PDF using a narrow prior centered on the true value of β.

Table 2 Conditional MSE (averaged over 1000 independent runs) for the Gamma example and the three fusion methods considered when $N \in \{10^3, 10^4, 10^5, 10^6, 5 \cdot 10^6\}$, $L \in \{1, 2, 5, 10, 25, 100, 200, 500, 1000\}$, and $N_\ell = N/L \in \{6, 12, 30, 60, 240, 600, 1200, 3000, 6000\}$.

Experiment		N_ℓ								
N	**Estimator**	5	10	20	50	100	200	500	100	N
10^3	EWF	0.3480	0.1011	0.0369	0.0193	0.0159	0.0148	0.0143	0.0143	0.0143
	SCMSE	0.2042	0.0637	0.0243	0.0147	0.0141	0.0142	0.0143	0.0143	
	Average	0.0191	0.0152	0.0144	0.0143	0.0143	0.0143	0.0143	0.0143	
10^4	EWF	0.3067	0.0695	0.0172	0.0035	0.0017	0.0013	0.0012	0.0012	0.0012
	SCMSE	0.2104	0.0598	0.0170	0.0038	0.0019	0.0014	0.0012	0.0012	
	Average	0.0034	0.0013	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	
10^5	EWF	0.3086	0.0695	0.0166	0.0027	0.0008	0.0003	0.0002	0.0001	0.0001
	SCMSE	0.2058	0.0566	0.0149	0.0025	0.0007	0.0003	0.0002	0.0001	
	Average	0.0027	0.0003	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	
10^6	EWF	0.3081	0.0691	0.0164	0.0025	0.0006	0.0002	0.000034	0.000016	0.000012
	SCMSE	0.2069	0.0568	0.0149	0.0025	0.0006	0.0002	0.000040	0.000020	
	Average	0.0025	0.0002	0.000020	0.000012	0.000012	0.000012	0.000012	0.000012	

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 the paper. We repeat the experiments 50 times and average the results. The results, shown in Table 5.2 and Figures 4–6, 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. Regarding the three scenarios considered, we note that the best performance is obtained in the second case (with $\mathrm{MSE}(\hat{\mathbf{x}}^{(\mathrm{LMSE})}|\mathbf{y}) = 0.0021$), 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.

Table 3 Conditional MSE (averaged over 50 independent runs) for the three scenarios and the four fusion methods considered when $N = 6000$, $L \in \{1, 2, 5, 10, 25, 100, 200, 500, 1000\}$, and $N_\ell = N/L \in \{6, 12, 30, 60, 240, 600, 1200, 3000, 6000\}$.

Experiment		$\boldsymbol{N_\ell}$								
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	

Finally, in order to study the scaling behaviour of the fusion rules as N increases, we have also simulated the three scenarios for $N = 30000$ as well as Scenario 2 for $N = 600000$. The results, displayed in Table 5.2 and Figure 7 respectively, show that the performance of all the fusion rules scales roughly as a function of the number of samples, N.

6 Conclusions

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

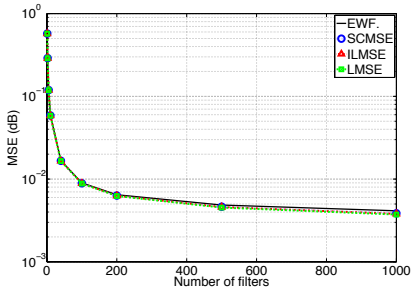

Figure 4 Localization Example. Conditional MSE as a function of L for Scenario 1 (Sc1) when $N = 6000$.

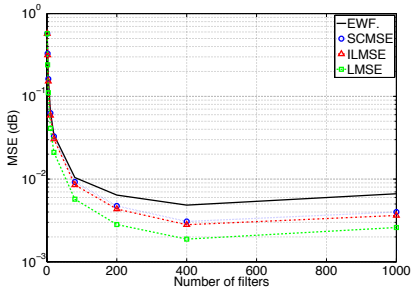

Figure 5 Localization Example. Conditional MSE as a function of L for Scenario 2 (Sc2) when $N = 6000$.

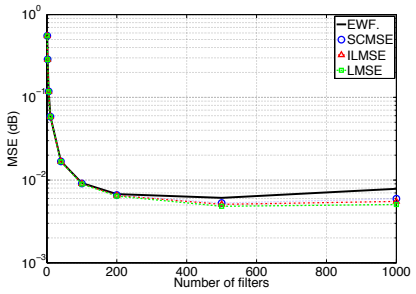

Figure 6 Localization Example. Conditional MSE as a function of L for Scenario 3 (Sc3) when $N = 6000$.

applying them to a simple problem where all the posterior densities followed a Gamma PDF, and 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. Furthermore, it has been shown that splitting the data can be advantageous in terms of attaining a reduced mean squared error (MSE), but only when the bias in the partial estimators can be controlled. In future works we plan to address bias correction approaches, as well as optimal linear fusion schemes for biased and/or correlated partial estimators, Some other interesting areas of research are non-linear fusion techniques and the development of fusion schemes

Table 4 Conditional MSE (averaged over 50 independent runs) for the three scenarios and the four fusion methods considered when $N = 30000$, $L \in \{1, 10, 25, 50, 125, 500, 1000, 2500, 5000\}$, and $N_\ell = N/L \in \{6, 12, 30, 60, 240, 600, 1200, 3000, 30000\}$.

Experiment		N_ℓ								
Scenario	**Estimator**	6	12	30	60	240	600	1200	3000	30000
Sc1	EWF	0.0008	0.001	0.0013	0.0018	0.0033	0.0117	0.0231	0.0574	0.5879
	SCMSE	0.0008	0.0009	0.0013	0.0018	0.0033	0.0117	0.023	0.0573	
	ILMSE	0.0008	0.0009	0.0013	0.0018	0.0033	0.0117	0.023	0.0572	
	LMSE	0.0007	0.0009	0.0012	0.0017	0.0033	0.0116	0.0229	0.057	
Sc2	EWF	0.0007	0.0009	0.0012	0.0018	0.0036	0.0131	0.0335	0.0661	0.5879
	SCMSE	0.0004	0.0006	0.0009	0.0015	0.0033	0.0125	0.0323	0.0638	
	ILMSE	0.0004	0.0005	0.0009	0.0014	0.0031	0.0118	0.0304	0.0611	
	LMSE	0.0003	0.0003	0.0006	0.0009	0.0021	0.0082	0.0214	0.0533	
Sc3	EWF	0.0018	0.0011	0.0013	0.0018	0.0033	0.0118	0.0229	0.0579	0.5751
	SCMSE	0.0014	0.001	0.0013	0.0018	0.0033	0.0118	0.0228	0.0577	
	ILMSE	0.0013	0.001	0.0013	0.0018	0.0033	0.0118	0.0228	0.0576	
	LMSE	0.0012	0.001	0.0013	0.0018	0.0033	0.0117	0.0227	0.0574	

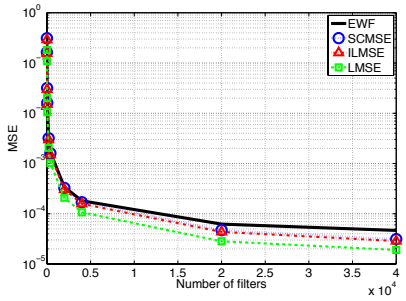

Figure 7 Localization Example. Conditional MSE as a function of L for Scenario 3 (Sc2) when $N = 600000$.

where the partial Monte Carlo estimators are allowed to exchange a reduced amount of information.

Acknowledgements
This work has been supported by the Spanish government's projects DISSECT (TEC2012-38058-C03-01), AGES (S2010/BMD-2422), ALCIT (TEC2012-38800-C03-01), COMPREHENSION (TEC2012-38883-C02-01), and OTOSiS (TEC2013-41718-R); by the BBVA Foundation through project MG-FIAR ("I Convocatoria de Ayudas Fundación BBVA a Investigadores, Innovadores y Creadores Culturales"); by the National Science Foundation under Award CCF-0953316; and by the European Union's 7th FP through the Marie Curie ITN MLPM2012 (Grant No. 316861).

Author details
[1]Dep. of Signal Theory and Communications, Technical Univ. of Madrid, Madrid, Spain. [2]Institute of Mathematical Sciences and Computing, University of Sao Paulo, Sao Carlos, Brazil. [3]Dep. of Signal Theory and Communications, Univ. Carlos III of Madrid, Madrid, Spain. [4]Dep. of Electrical and Computer Eng., Stony Brook Univ., New York, USA.

References

1. Van Trees, H.L.: Detection, Estimation and Modulation Theory. John Wiley and Sons, Hoboken, NJ (USA) (1968)
2. Casella, G., Berger, R.L.: Statistical Inference. Duxbury, ??? (2002)
3. Scharf, L.L.: Statistical Signal Processing. Addison-Wesley, Reading, MA (USA) (1991)

4. Kay, S.M.: Fundamentals of Statistical Signal Processing: Estimation Theory. Prentice Hall, Upper Saddle River, NJ (USA) (1993)
5. Mendel, J.M.: Lessons in Estimation Theory for Signal Processing, Communications, and Control. Pearson Education, New York, NY (USA) (1995)
6. Rasmussen, C.E.: Gaussian processes for machine learning (2006)
7. Hjort, N.L., Holmes, C., Müller, P., Walker, S.G.: Bayesian Nonparametrics vol. 28. Cambridge University Press, Cambridge (UK) (2010)
8. Gibbons, J.D., Chakraborti, S.: Nonparametric Statistical Inference. Springer, ??? (2011)
9. Giannakis, G.B., Bach, F., Cendrillon, R., Mahoney, M., Neville, J.: Signal processing for big data (special issue). IEEE Signal Processing Magazine **31**(5), 15–111 (2014)
10. Djuric, P.M., Goodsill, S. (eds.): Special issue on Monte Carlo methods for statistical signal processing. IEEE Transactions on Signal Processing **50**(2), 173–173 (2002)
11. Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer, ??? (2004)
12. Doucet, A., Wang, X.: Monte Carlo methods for signal processing: a review in the statistical signal processing context. Signal Processing Magazine, IEEE **22**(6), 152–170 (2005)
13. Chopin, N.: A sequential particle filter method for static models. Biometrika **89**(3), 539–552 (2002)
14. Scott, S.L., Blocker, A.W., Bonassi, F.V., Chipman, H.A., George, E.I., McCulloch, R.E.: Bayes and big data: The consensus Monte Carlo algorithm. In: EFaBBayes 250th Conference, vol. 16 (2013)
15. Suchard, M.A., Wang, Q., Chan, C., Frelinger, J., Cron, A., West, M.: Understanding gpu programming for statistical computation: Studies in massively parallel massive mixtures. Journal of Computational and Graphical Statistics **19**(2), 419–438 (2010)
16. Lee, A., Yau, C., Giles, M.B., Doucet, A., Holmes, C.C.: On the utility of graphics cards to perform massively parallel simulation of advanced monte carlo methods. Journal of computational and graphical statistics **19**(4), 769–789 (2010)
17. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Communications of the ACM **51**(1), 107–113 (2008)
18. Wallis, K.F.: Combining forecasts – forty years later. Applied Financial Economics **21**(1–2), 33–41 (2011)
19. Bates, J.M., Granger, C.W.: The combination of forecasts. Operational Research Quarterly **20**(4), 451–468 (1969)
20. Dickinson, J.P.: Some statistical results in the combination of forecasts. Operational Research Quarterly **24**, 253–260 (1975)
21. Bordley, R.F.: The combination of forecasts: A Bayesian approach. Journal of the Operational Research Society **33**(2), 171–174 (1982)
22. Lavancier, F., Rochet, P.: A general procedure to combine estimators. arXiv preprint arXiv:1401.6371 (2014)
23. Allenby, G.M.: Cross-validation, the bayes theorem, and small-sample bias. Journal of Business & Economic Statistics **8**(2), 171–178 (1990)
24. Predd, J.B., Kulkarni, S.R., Poor, H.V.: Distributed learning in wireless sensor networks. IEEE Signal Processing Magazine **23**(4), 56–69 (2006)
25. Xiao, J.-J., Ribeiro, A., Luo, Z.-Q., Giannakis, G.B.: Distributed compression-estimation using wireless sensor networks. IEEE Signal Processing Magazine **23**(4), 27–41 (2006)
26. Swami, A., Zhao, Q., Hong, Y.-W., Tong, L. (eds.): Wireless Sensor Networks: Signal Processing and Communications Perspectives. John Wiley and Sons, ??? (2007)
27. Cetin, M., Chen, L., III, J.W.F., Ihler, A.T., Moses, R.L., Wainwright, M.J., Willsky, A.S.: Distributed fusion in sensor networks. IEEE Signal Processing Magazine **23**(4), 42–55 (2006)
28. Olfati-Saber, R., Fax, J.A., Murray, R.M.: Consensus and cooperation in networked multi-agent systems. Proceedings of the IEEE **95**(1), 215–233 (2007)
29. Dimakis, A.G., Kar, S., Moura, J.F., Rabbat, M.G., Scaglione, A.: Gossip algorithms for distributed signal processing. Proceedings of the IEEE **98**(11), 1847–1864 (2010)
30. Cattivelli, F.S., Sayed, A.H.: Diffusion LMS strategies for distributed estimation. IEEE Transactions on Signal Processing **58**(3), 1035–1048 (2010)
31. Wilkinson, D.J.: Parallel Bayesian computation. Statistics Textbooks and Monographs **184** (2006)
32. Wilkinson, B., Allen, M.: Parallel programming: techniques and applications using networked workstations and parallel computers (1998)
33. Huang, Z., Gelman, A.: Sampling for bayesian computation with large datasets. Available at SSRN 1010107 (2005)
34. Neiswanger, W., Wang, C., Xing, E.: Asymptotically exact, embarrassingly parallel MCMC. arXiv:1311.4780v2, 1–16 (21 Mar. 2014)
35. Wang, X., Dunson, D.B.: Parallelizing MCMC via Weierstrass sampler. arXiv:1312.4605v2 (25 May 2014)
36. Luengo, D., Martino, L., Elvira, V., Bugallo, M.: Efficient linear combination of partial Monte Carlo estimators. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2015)
37. Elvira, V., Martino, L., Luengo, D., Bugallo, M.F.: Efficient multiple importance sampling estimators. Signal Processing Letters, IEEE **22**(10), 1757–1761 (2015)
38. Le Cam, L.: Asymptotic Methods in Statistical Decision Theory. Springer, New York, NY (USA) (1986)
39. van der Vaart, A.W.: Asymptotic Statistics. Cambridge University Press, ??? (1998)