CERFACS

Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique

Multivariate extensions of the Multilevel Best Linear Unbiased Estimator for ensemble-variational data assimilation

Mayeul Destouches, Paul Mycek, Selime Gürol

For comments and reactions, please email
mayeul.destouches@umr-cnrm.fr

Technical Report TR-PA-23-67

1 June 2023

Publications of the Parallel Algorithms Team

https://cerfacs.fr/publication/

Abstract

Multilevel estimators aim at reducing the variance of Monte Carlo statistical estimators, by combining samples generated with simulators of different costs and accuracies. In particular, the recent work of Schaden and Ullmann (2020) on the multilevel best linear unbiased estimator (MLBLUE) introduces a framework unifying several multilevel and multifidelity techniques. The MLBLUE is reintroduced here using a variance minimization approach rather than the regression approach of Schaden and Ullmann. We then discuss possible extensions of the scalar MLBLUE to a multidimensional setting, i.e. from the expectation of *scalar* random variables to the expectation of random *vectors*. Several estimators of increasing complexity are proposed: a) multilevel estimators with scalar weights, b) with element-wise weights, c) with spectral weights and d) with general matrix weights. The computational cost of each method is discussed. We finally extend the MLBLUE to the estimation of second-order moments in the multidimensional case, i.e. to the estimation of covariance matrices. The multilevel estimators proposed are d) a multilevel estimator with scalar weights and e) with element-wise weights. In large-dimension applications such as data assimilation for geosciences, the latter estimator is computationnally unaffordable. As a remedy, we also propose f) a multilevel covariance matrix estimator with optimal multilevel localization, inspired by the optimal localization theory of Ménétrier and Auligné (2015). Some practical details on weighted MLMC estimators of covariance matrices are given in appendix.

Contents

1 Introduction

Multilevel techniques aim at reducing the variance of Monte Carlo statistical estimators, typically for the estimation of the expectation of a scalar random variable. These techniques combine in an astute way samples obtained through numerical simulators of varying accuracy and cost. An example of popular multilevel technique is the Multilevel Monte Carlo (MLMC) method, popularized by Giles (2008, 2015).

Recently, an interesting unifying framework was proposed by Schaden and Ullmann (2020, 2021), herafter SU20 and SU21. Among others, the framework of Schaden and Ullmann includes multilevel Monte Carlo techniques (MLMC, Giles, 2008, 2015 for a review), multifidelity techniques (Peherstorfer et al., 2018) and approximate control variates (Gorodetsky et al., 2020). In this unified framework, some (new) estimators naturally appear as optimal, the so-called *Multilevel Best Linear Unbiased Estimators*, MLBLUEs. This framework has been complemented by Croci et al. (2023), who propose an efficient algorithm to solve the *model selection and sample allocation problem* (MOSAP) for the MLBLUE.

The present note proposes a new way to derive the MLBLUE, by building a weighted multilevel estimator and optimizing its weights to minimize the estimator's variance under a no-bias constraint. It also gives some insight on how the MLBLUE approach can be extended to the estimation of first and second-order statistical moments of random vectors, in possibly large dimensions.

This extension to second-order moments and to random vectors is motivated by possible applications in ensemble-variational data assimilation, where Monte Carlo methods are used at a key stage, to estimate the covariance matrix of forecast errors (Lorenc, 2003; Buehner, 2005 and Bannister, 2017 for a review). As a result, the present note is not as general as the original articles by SU20 and SU21, nor as mathematically grounded. The authors are biased towards MLMC-like applications, and towards the estimation of discrete covariance operators in large dimension for geoscience applications.

The note is organized as follows. Section 2 presents the main results of SU20 and SU21. Section 3 presents another way to derive these results, based on direct minimization of the variance of a weighted multilevel estimator. The next sections proposes extensions of the MLBLUE, some of which are unpublished in the literature to the best of our knowledge. We propose an extension to the multi-dimensional case (estimation of the expectation of a random vector) in section 4. We propose an extension to the estimation of covariance and covariance matrices in sections 5 and 6, including an extension to optimal localization for multilevel covariance matrices in the line of Ménétrier et al. (2015a,b), hereafter M15a and M15b.

2 The MLBLUE: reminder and notations

Let $Z_\ell = f_\ell(X)\colon \Omega \to \mathbb{R}$, $1 \leq \ell \leq L$ be a set of random variables approximating $f_L(X)$, where f_L is a costly numerical simulator. The ℓ indexing the f_ℓ models are hereafter called fidelity levels. These fidelity levels may be associated to different spatial meshes, from the coarsest to the finest ($\ell = 1$ to $\ell = L$ for an MLMC-like structure). This is not required though, and what follows can be applied even if the fidelities come from other sources, and even if there is no clear ranking of the models according to their accuracy.

Example The hierarchy of simulators can be forecasting models $f_\ell\colon \mathbb{R}^n \to \mathbb{R}$ running on meshes with finer and finer horizontal resolutions, and predicting temperature at one given location. $X\colon \Omega \to \mathbb{R}^n$ can be a random vector representing uncertain initial conditions of a numerical weather forecast. We are interested in the mean temperature that is forecast by the finest model, $\mathbb{E}\big[f_L(X)\big]$.

Multilevel techniques rely on coupled simulations, i.e. simulations that run at different levels using the same stochastic input X. The sets of coupled levels can be sorted in K coupling groups $(S^{(k)})_{k=1}^K$. More formally, let $(S^{(k)})_{k=1}^K$ be a family of subsets of $\{1, \dots, L\}$. We impose

$$S^{(k)} \neq S^{(k')} \text{ for } k \neq k' \tag{1}$$

$$\cup_{k=1}^K S^{(k)} = \{1, \dots, L\}. \tag{2}$$

We denote by $p^{(k)}$ the cardinality of $S^{(k)}$.

For $1 \leq k \leq K$, $R^{(k)}\colon \mathbb{R}^L \to \mathbb{R}^{p^{(k)}}$ is the selection operator for group $S^{(k)}$ verifying $\forall x \in \mathbb{R}^L$, $R^{(k)}x = (x_\ell)_{\ell \in S^{(k)}}$. The associated extension operator is $P^{(k)} := \big(R^{(k)}\big)^\intercal\colon \mathbb{R}^{p^{(k)}} \to \mathbb{R}^L$.

MLMC-like example We can use this formalism to describe the coupling structure of an MLMC estimator with three levels, as is done in example 2.1 of SU20. The coupling groups in this case are $S^{(1)} = \{1\}$, $S^{(2)} = \{1, 2\}$ and $S^{(3)} = \{2, 3\}$. The associated extension operators are

$$P^{(1)} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \tag{3}$$

$$P^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{4}$$

$$P^{(3)} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{5}$$

The repartition of simulators among coupling groups $S^{(k)}$ can be visualized using the tableaux used by Schaden and Ullmann, and reproduced for this example in figure 1c.

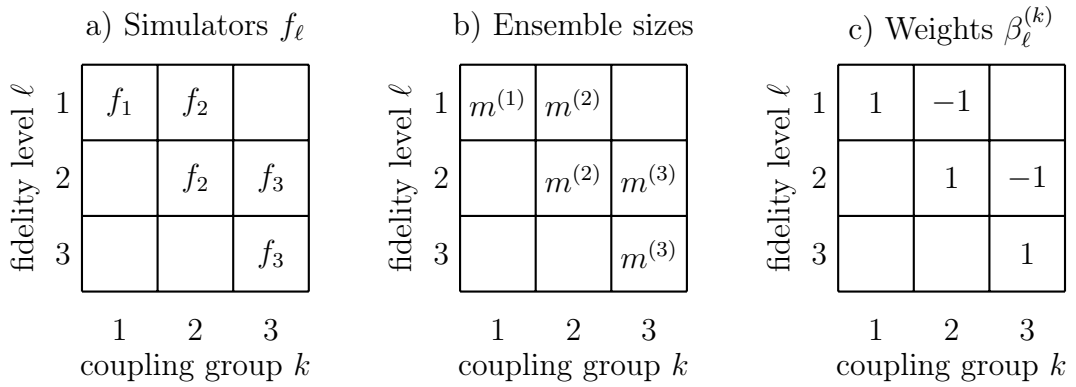

Figure 1: MLMC coupling structure. Tableaux inspired by SU20.

We denote by $\mu := (\mathbb{E}[Z_\ell])_{\ell=1}^L$ the vector of expectations at each fidelity level. We are interested in estimating $\alpha^\intercal \mu$ for a given vector $\alpha \in \mathbb{R}^L \setminus \{0\}$. In practice, $\alpha = e_L := (0, \cdots, 0, 1)^\intercal$, but this is not mandatory.

Let $m^{(1)}, \ldots, m^{(K)}$ be the number of available simulations for each group k (see figure 1b). SU20 provide the best estimator for $\alpha^\intercal \mu$ among the unbiased estimators that linearly combine the simulations $f_\ell(X^{(k,i)})$ for $1 \leq k \leq K$, $\ell \in S^{(k)}$ and $1 \leq i \leq m^{(k)}$, where the $X^{(k,i)}$ are i.i.d. random variables following the same law as X. These linear estimators are of the form

$$\widehat{\mu}^{\mathrm{ML}} := \sum_{k=1}^{K} \sum_{\ell \in S^{(k)}} \beta_\ell^{(k)} \widehat{E}^{(k)}[Z_\ell] \tag{6}$$

where $\widehat{E}^{(k)}[Z_\ell]$ is the standard Monte Carlo estimator for $\mathbb{E}[Z_\ell]$, using $m^{(k)}$ random inputs associated to group k,

$$\widehat{E}^{(k)}[Z_\ell] := \frac{1}{m^{(k)}} \sum_{i=1}^{m^{(k)}} f_\ell(X^{(k,i)}). \tag{7}$$

The inner sum in (6) can be written as a scalar product by introducing two more notations. Firstly, we denote by $\beta^{(k)} := \left(\beta_\ell^{(k)}\right)_{\ell \in S^{(k)}} \in \mathbb{R}^{p^{(k)}}$ the vector of weights associated to group k. In the case of a three-level MLMC, we would have the (sub-optimal) weights $\beta^{(1)} = (1)$ and $\beta^{(2)} = \beta^{(3)} = (-1, 1)^\intercal$ (figure 1c).

Secondly, we denote by $Z^{(k)} := (Z_\ell)_{\ell \in S^{(k)}}$ the random vector gathering all random variables in group k. Then equation (6) becomes

$$\widehat{\mu}^{\mathrm{ML}} = \sum_{k=1}^{K} \left(\beta^{(k)}\right)^{\intercal} \widehat{E}^{(k)}\left[Z^{(k)}\right] \tag{8}$$

The (optimal) vector $\beta^{(k)}$ can be expressed as (from equation 2.7 in SU21)

$$\beta^{(k)} = m^{(k)} \left(C^{(k)}\right)^{-1} R^{(k)} \left(\sum_{k'=1}^{K} m^{(k')} P^{(k')} \left(C^{(k')}\right)^{-1} R^{(k')} \right)^{-1} \alpha, \tag{9}$$

where $C^{(k)} := \mathrm{Cov}(Z^{(k)}, Z^{(k)})$ and $\mathrm{Cov}(A, B) := \mathbb{E}[(A - \mathbb{E}[A])\,(B - \mathbb{E}[B])^{\intercal}]$ denotes the covariance matrix of random vectors A and B. These $C^{(k)}$ matrices are unknown in practice and must be estimated, which results in sub-optimal weights. Note that the estimation of the $C^{(k)}$ should be done independently of the estimation of $\alpha^{\intercal}\mu$, otherwise a bias is introduced. The importance of this bias is likely to depend on the particular application and setting considered.

Model selection and sample allocation problem This approach provides the MLBLUE for a given coupling structure and a given number of samples on each coupling group. SU20 propose some ways to optimize the model selection and sample allocation by minimizing the variance of the associated MLBLUE. Their approach has been later extended and made more robust by Croci et al. (2023), who transform it into a semidefinite programming problem.

3 Retrieving the MLBLUE via variance minimization

We believe the derivation based on constrained minimization of the variance to be more direct, and perhaps more intuitive than the regression approach proposed by SU20. Though both derivations are closely related, we believe the variance minimization approach may appear as more natural to some readers, especially from the community of multifidelity estimation methods based on control variates (see for instance Gorodetsky et al., 2020). We derive this approach here.

Given the fidelity levels $1, \dots, L$ and the coupling structure $(S^{(k)}, m^{(k)})_{k=1}^K$, we look for an unbiased estimator of $\alpha^\intercal \mu$ that linearly combines the samples and that has the lowest possible variance (the BLUE). We assume that the multilevel estimator is of the form of equations (6) and (8), and we look for the $\beta_\ell^{(k)}$ coefficients that minimize the variance under a no-bias constraint.

Unbiasedness constraint The optimal β weights are subject to the unbiasedness constraint

$$\mathbb{E}[\widehat{\mu}(\beta)] = \alpha^\intercal \mu \quad \iff \quad \mathbb{E}\left[\sum_{k=1}^K \left(\beta^{(k)}\right)^\intercal \widehat{E}^{(k)}\left[Z^{(k)}\right]\right] = \alpha^\intercal \mu \tag{10}$$

$$\iff \quad \sum_{k=1}^K \left(\beta^{(k)}\right)^\intercal R^{(k)} \mu = \alpha^\intercal \mu \tag{11}$$

$$\iff \quad \sum_{k=1}^K P^{(k)} \beta^{(k)} = \alpha, \tag{12}$$

assuming that we have no prior information on μ, so that the unbiasedness should be met for all values of $\mu \in \mathbb{R}^L$.

$$\mathbb{E}[\widehat{\mu}(\beta)] = \alpha^\intercal \mu \quad \iff \quad \begin{pmatrix} P^{(1)} & \cdots & P^{(K)} \end{pmatrix} \beta = \alpha \tag{13}$$

$$\iff \quad g(\beta) = 0 \tag{14}$$

$$\text{with } g(\beta) := \begin{pmatrix} P^{(1)} & \cdots & P^{(K)} \end{pmatrix} \beta - \alpha \tag{15}$$

and where we denote by $\beta := \left(\beta^{(k)}\right)_{k=1}^K \in \mathbb{R}^p$ the vector made of all the $\beta^{(k)}$, with $p := \sum_{k=1}^K p^{(k)}$.

Expression of the variance The variance of the linear estimator $\widehat{\mu}$ is the sum of the variances for each coupling group, since simulations are independent from

one coupling group to another. We denote by $\mathbb{V}(X)$ the variance of any square-integrable random variable X.

$$\mathbb{V}(\widehat{\mu}(\beta)) = \sum_{k=1}^{K} \mathbb{V}\Big(\big(\beta^{(k)}\big)^\intercal \, \widehat{E}^{(k)}\big[Z^{(k)}\big] \Big) \tag{16}$$

$$= \sum_{k=1}^{K} \big(\beta^{(k)}\big)^\intercal \operatorname{Cov}\Big(\widehat{E}^{(k)}\big[Z^{(k)}\big], \widehat{E}^{(k)}\big[Z^{(k)}\big] \Big) \, \beta^{(k)} \tag{17}$$

$$= \sum_{k=1}^{K} \frac{1}{m^{(k)}} \big(\beta^{(k)}\big)^\intercal \operatorname{Cov}\big(Z^{(k)}, Z^{(k)}\big) \, \beta^{(k)} \tag{18}$$

$$= \sum_{k=1}^{K} \frac{1}{m^{(k)}} \big(\beta^{(k)}\big)^\intercal C^{(k)} \beta^{(k)} \tag{19}$$

$$= \beta^\intercal \Sigma \beta \tag{20}$$

$$\text{with } \Sigma := \operatorname{Diag}_{k=1}^{K}\left(\frac{1}{m^{(k)}} C^{(k)} \right). \tag{21}$$

We used the independence of the $f_\ell\big(X^{(k,i)}\big)$ for different i to go from (17) to (18), and used the Diag operator to denote a block-diagonal matrix. Equation (19) is equivalent to equation (2.8) in SU21.

Convexity of the variance Σ is a block-diagonal matrix with covariance matrices on the diagonal. As a result, it is positive semi-definite and $\mathbb{V}(\widehat{\mu}(\beta))$ is a convex (quadratic) function of β. Note that as will be discussed hereafter, the covariance matrices $C^{(k)}$ are actually positive definite, and the variance is a strictly convex function of β.

Constrained minimization problem The best unbiased estimator is then given by the minimizer of the variance under the unbiasedness constraint.

$$\beta^\star = \underset{\beta \text{ s.t. } g(\beta)=0}{\arg\min} \frac{1}{2} \mathbb{V}\big(\widehat{\mu}(\beta)\big). \tag{22}$$

Unconstrained minimization problem The assumptions on the coupling groups $S^{(k)}$ ensure that the L constraints are linearly independent. A vector $\lambda \in \mathbb{R}^L$ of Lagrange multipliers can be used to solve the minimization problem. From the

convexity of the quadratic problem, the solutions of (22) are the solutions of

$$\beta^\star, \lambda^\star = \arg\min_{\beta,\lambda} \mathcal{L}(\beta, \lambda), \tag{23}$$

$$\text{with } \mathcal{L}(\beta, \lambda) = \frac{1}{2}\mathbb{V}(\widehat{\mu}(\beta)) - \lambda^\intercal g(\beta). \tag{24}$$

In particular, the gradient of the Lagrangian,

$$\nabla_\beta \mathcal{L} = \beta^\intercal \operatorname{Diag}_{k=1}^{K}\left(\frac{1}{m^{(k)}} C^{(k)}\right) - \lambda^\intercal \begin{pmatrix} P^{(1)} & \cdots & P^{(K)} \end{pmatrix}, \tag{25}$$

$$\nabla_\lambda \mathcal{L} = \beta^\intercal \begin{pmatrix} P^{(1)} & \cdots & P^{(K)} \end{pmatrix}^\intercal - \alpha^\intercal, \tag{26}$$

should vanish. The associated linear system is

$$\left(\begin{array}{ccc|c} \frac{1}{m^{(1)}} C^{(1)} & & & -R^{(1)} \\ & \ddots & & \vdots \\ & & \frac{1}{m^{(k)}} C^{(K)} & -R^{(K)} \\ \hline P^{(1)} & \cdots & P^{(K)} & 0_L \end{array}\right) \left(\begin{array}{c} \beta^\star \\ \hline \lambda^\star \end{array}\right) = \left(\begin{array}{c} 0 \\ \vdots \\ 0 \\ \hline \alpha \end{array}\right) \tag{27}$$

To simplify the notations, we drop the stars and write the system as

$$\Sigma\beta - P^\intercal\lambda = 0, \tag{28}$$

$$P\beta = \alpha. \tag{29}$$

This system can be solved by substitution:

$$\text{From (28):} \qquad \beta = \Sigma^{-1}P^\intercal\lambda \tag{30}$$

$$\text{Inserting (30) in (29):} \qquad P\Sigma^{-1}P^\intercal\lambda = \alpha \tag{31}$$

$$\iff \lambda = \left(P\Sigma^{-1}P^\intercal\right)^{-1}\alpha \tag{32}$$

$$\text{Inserting (32) in (30):} \qquad \beta = \Sigma^{-1}P^\intercal\left(P\Sigma^{-1}P^\intercal\right)^{-1}\alpha. \tag{33}$$

Invertibility of the matrices We assumed the invertibility of Σ and $P\Sigma^{-1}P^\intercal$. The invertibility of Σ follows from the invertibility of each covariance matrix $C^{(k)}$. Suppose Σ is singular. Then, there exists a k such that $C^{(k)}$ is singular. Then there exists a vector $\gamma \in \mathbb{R}^{p^{(k)}} \setminus \{0\}$ such that $\gamma^\intercal C^{(k)}\gamma = \mathbb{V}(\gamma^\intercal Z^{(k)}) = 0$, *i.e.* $\sum_{\ell \in S^{(k)}} \gamma_\ell Z_\ell$ is actually deterministic. One random variable Z_ℓ can thus be expressed as an affine function of the others. It brings no new information to the problem, and can be removed from the estimator. The invertibility of $P\Sigma^{-1}P^\intercal$ follows from the invertibility of Σ^{-1} and from P being a full-rank linear map from $\mathbb{R}^p$ to the lower-dimensional space $\mathbb{R}^L$.

MLBLUE weights The optimal choice of β is thus given by equation (33):

$$\beta = \begin{pmatrix} \beta^{(1)} \\ \vdots \\ \beta^{(K)} \end{pmatrix} = \begin{pmatrix} m_1 \left(C^{(1)}\right)^{-1} & & \\ & \ddots & \\ & & m^{(k)} \left(C^{(K)}\right)^{-1} \end{pmatrix} \begin{pmatrix} R^{(1)} \\ \vdots \\ R^{K} \end{pmatrix} \phi^{-1}\alpha, \tag{34}$$

where

$$\phi := P\Sigma^{-1}P^{\intercal} = \sum_{k=1}^{K} m^{(k)} P^{(k)} \left(C^{(k)}\right)^{-1} R^{(k)}. \tag{35}$$

For a given group k, we retrieve equation (2.7) of SU21, namely

$$\beta^{(k)} = m^{(k)} \left(C^{(k)}\right)^{-1} R^{(k)} \left(\sum_{k'=1}^{K} m^{(k')} P^{(k')} \left(C^{(k')}\right)^{-1} R^{(k')} \right)^{-1} \alpha. \tag{36}$$

Is it the MLBLUE? The estimators of the form (8) only describe a specific subset of all possible linear estimators. The more general class of linear estimators would be

$$\mu^{\mathrm{ML}} = \sum_{k=1}^{K} \sum_{\ell \in S^{(k)}} \sum_{i=1}^{m^{(k)}} \beta_\ell^{(k,i)} f_\ell\left(X^{(k,i)}\right). \tag{37}$$

Estimators of the form (37) can be related to Eq. (8) by replacing $\beta_\ell^{(k,i)}$ with $\beta_\ell^{(k)}/m^{(k)}$. In other words, Eq. (8) assumes that at the optimum, the $\beta_\ell^{(k,i)}$ weights should not depend on the sample index i. This independence is a very intuitive result, that can be derived from the interchangeability of the samples i and the strict convexity of the variance of (37) as a function of the weights.

Sample allocation Inserting (36) into (19) and simplifying gives the expression of the minimum variance reachable with a given sample allocation m (equivalent to equation 2.12 in SU20):

$$\mathbb{V}(\mu^{\mathrm{ML}}(m)) = \alpha^{\intercal}\phi(m)^{-1}\alpha. \tag{38}$$

The optimal choice for the sample allocation $m = (m^{(1)}, \ldots, m^{(K)})$ is given by minimizing this variance under a computational cost constraint, which can be done numerically.

Model selection and sample allocation problem Alternatively, the model selection and sample allocation problem (MOSAP) can be solved through a semidefinite programming problem, as shown by Croci et al. (2023), in the typical case where $\alpha = e_L$:

$$\min_{m \geq 0, t} t \quad \text{s.t.} \quad \begin{cases} \begin{pmatrix} \phi(m) & e_L \\ e_L^\intercal & t \end{pmatrix} \text{ is positive semi-definite,} \\ m^\intercal c \leq b, \\ m^\intercal h \geq 1. \end{cases} \tag{39}$$

The second constraint imposes a computational budget b, where $c = (c^{(1)}, \dots, c^{(K)})^\intercal$ describes the computational cost of generating a coupled sample in each coupling group. The third constraint, where h denotes the vector of $\{0;1\}^K$ such that $h^{(k)} = 1$ if and only if $L \in S^{(k)}$, enforces that the high-fidelity model be sampled at least once.

Note that this extends easily to the case of any α by replacing e_L with α and by adding inequality constraints to ensure that all model with non-zero coefficients in α are sampled at least once. Also note that this approach to the MOSAP can handle sample sizes of zero, which means the problem can be directly optimized on the set of all possible coupling groups. This would not be directly possible with the sample allocation strategy proposed previously.

A similar version for a target accuracy with no constraint on the computational budget is also proposed in Croci et al. (2023).

4 Estimation of the expectation of a random vector

This section extends the MLBLUE methodology to the estimation of the expectation of a random vector. Section (4.1) introduces new notations to deal with vector quantities. We then propose various multilevel estimators of increasing complexity (sections 4.2 – 4.4), before introducing the general multidimensional MLBLUE in section 4.5.

4.1 Notations for the multidimensional case

We consider the case where $\mathbf{Z}_\ell : \Omega \to \mathbb{R}^n$ are random vectors. All vectors or matrices related to multidimensional quantities are written in bold.

We are interested in the expectation $\boldsymbol{\mu} := \mathbb{E}\big[\mathbf{Z}\big]$, where $\mathbf{Z} := (\mathbf{Z}_1 \dots \mathbf{Z}_L)^\intercal$ is the random matrix with values in $\mathbb{R}^{L\times n}$, obtained by stacking the random vectors from all fidelity levels so that the first dimension of $\mathbf{Z}$ indexes the fidelity levels.

We want to estimate a linear combination of the expectations on different levels, $\boldsymbol{\mu}_\alpha := \sum_{\ell=1}^L \alpha_\ell \, \mathbb{E}\big[\mathbf{Z}_\ell\big] = \boldsymbol{\mu}^\intercal \alpha$, where α is a non-zero vector of $\mathbb{R}^L$. In practice, we are often interested in the estimation for one given fidelity level, typically the highest, in which case $\alpha = e_L$.

The selection and extension operators from the scalar case naturally extend to the vector case:

$$\mathbf{Z}^{(k)} := R^{(k)}\mathbf{Z} \in \mathbb{R}^{p^{(k)}\times n}, \quad \forall\, 1 \le k \le K. \tag{40}$$

Vector equivalent of the variance Under a no-bias constraint, the variance of a random variable is its (scalar) mean squared error (MSE). In the multidimensional case, minimizing the MSE of an estimator $\widehat{\boldsymbol{\mu}}$ using the 2-norm is equivalent to minimizing the sum of scalar MSEs for all vector elements.

$$\mathbb{E}\big[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2\big] = \sum_{i=1}^n \mathbb{E}\big[(\widehat{\mu}_i - \mu_i)^2\big] \tag{41}$$

$$= \sum_{i=1}^n \mathbb{V}(\widehat{\mu}_i) \tag{42}$$

$$= \operatorname{Tr}\operatorname{Cov}\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\mu}}\right) \tag{43}$$

where Tr is the trace operator, and where we used the unbiasedness of the estimator $\widehat{\boldsymbol{\mu}}$. It can be seen from here that the natural generalization of the variance for a random vector is the trace of the covariance matrix, *i.e.* the sum of the variances of each vector element.

Each of the following sections introduces a class of estimators $\widehat{\boldsymbol{\mu}}(\boldsymbol{\beta})$ where $\boldsymbol{\beta}$ is a set of weights. Similarly to the scalar case, the optimal value $\boldsymbol{\beta}^\star$ of these weights is found by minimization of the trace of the covariance matrix of the estimator, under a no-bias constraint:

$$\boldsymbol{\beta}^\star = \underset{\boldsymbol{\beta} \text{ s.t. } \mathbb{E}[\widehat{\boldsymbol{\mu}}(\boldsymbol{\beta})]=0}{\arg\min} \operatorname{Tr}\operatorname{Cov}\big(\widehat{\boldsymbol{\mu}}(\boldsymbol{\beta}), \widehat{\boldsymbol{\mu}}(\boldsymbol{\beta})\big). \tag{44}$$

Hereafter, the term *variance* is sometimes used to refer to the trace of the covariance matrix.

4.2 Scalar weights

The simplest possibility is to use scalar weights $\beta_\ell^{(k)}$, common to all random vector elements. In this case, $\boldsymbol{\mu}_\alpha$ is estimated through the linear combination of Monte Carlo estimators

$$\widehat{\boldsymbol{\mu}}_\alpha^{\text{sw}} = \sum_{k=1}^{K} \sum_{\ell \in S^{(k)}} \beta_\ell^{(k)} \widehat{E}^{(k)}\big[\mathbf{Z}_\ell\big] \tag{45}$$

$$= \sum_{k=1}^{K} \widehat{E}^{(k)}\big[\mathbf{Z}^{(k)}\big]^\intercal \beta^{(k)}. \tag{46}$$

The BLUE has no reason to lie in this class of estimators, which is just a subset of the linear estimators we will introduce in sections 4.3 to 4.5. Finding the optimal scalar weights $\beta_\ell^{(k)}$ is still of interest though, as these scalar-weigthed estimators are both simple and numerically affordable.

Unbiasedness constraint

$$\mathbb{E}\big[\widehat{\boldsymbol{\mu}}_\alpha^{\mathrm{sw}}\big] = \boldsymbol{\mu}_\alpha \tag{47}$$

$$\iff \quad \mathbb{E}\left[\sum_{k=1}^{K} \widehat{E}^{(k)}\big[\mathbf{Z}^{(k)}\big]^\intercal \beta^{(k)}\right] = \mathbb{E}[\mathbf{Z}^\intercal]\alpha \tag{48}$$

$$\iff \quad \sum_{k=1}^{K} \mathbb{E}\big[\mathbf{Z}^{(k)}\big]^\intercal \beta^{(k)} = \mathbb{E}[\mathbf{Z}^\intercal]\alpha \tag{49}$$

$$\iff \quad \sum_{k=1}^{K} \mathbb{E}\big[R^{(k)}\mathbf{Z}\big]^\intercal \beta^{(k)} = \mathbb{E}[\mathbf{Z}]^\intercal\alpha \tag{50}$$

$$\iff \quad \mathbb{E}\big[\mathbf{Z}\big]^\intercal \sum_{k=1}^{K} P^{(k)}\beta^{(k)} = \mathbb{E}[\mathbf{Z}]^\intercal\alpha \tag{51}$$

$$\iff \quad \sum_{k=1}^{K} P^{(k)}\beta^{(k)} = \alpha \tag{52}$$

$$\iff \quad g(\beta) = 0 \tag{53}$$

where g was defined in Eq. (15). This is exactly the unbiasedness constraint (14) from the scalar expectation case.

Variance of the estimator Herafter, the covariance operator is occasionally extended to random matrices using the definition

$$\mathrm{Cov}\big(\mathbf{A}, \mathbf{B}\big) := \mathbb{E}\big[\big(\mathbf{A} - \mathbb{E}[\mathbf{A}]\big)\big(\mathbf{B} - \mathbb{E}[\mathbf{B}]\big)^\intercal\big] \tag{54}$$

where $\mathbf{A}$ and $\mathbf{B}$ are any random matrices with same number of columns. This definition coincides with the usual one in the case of column vectors.

$$\operatorname{Tr}\operatorname{Cov}\left(\widehat{\boldsymbol{\mu}}_\alpha^{\mathrm{sw}}, \widehat{\boldsymbol{\mu}}_\alpha^{\mathrm{sw}}\right) = \sum_{k=1}^{K} \frac{1}{m^{(k)}} \operatorname{Tr}\operatorname{Cov}\left(\mathbf{Z}^{(k)\intercal}\beta^{(k)},\ \mathbf{Z}^{(k)\intercal}\beta^{(k)}\right) \tag{55}$$

$$= \sum_{k=1}^{K} \frac{1}{m^{(k)}} \operatorname{Tr}\operatorname{Cov}\left(\beta^{(k)\intercal}\mathbf{Z}^{(k)},\ \beta^{(k)\intercal}\mathbf{Z}^{(k)}\right) \tag{56}$$

$$= \sum_{k=1}^{K} \frac{1}{m^{(k)}} \operatorname{Tr}\left\{\beta^{(k)\intercal} \operatorname{Cov}\left(\mathbf{Z}^{(k)}, \mathbf{Z}^{(k)}\right)\beta^{(k)}\right\} \tag{57}$$

$$= \sum_{k=1}^{K} \frac{1}{m^{(k)}} \beta^{(k)\intercal} \operatorname{Cov}\left(\mathbf{Z}^{(k)}, \mathbf{Z}^{(k)}\right)\beta^{(k)} \tag{58}$$

$$= \sum_{k=1}^{K} \frac{1}{m^{(k)}} \beta^{(k)\intercal} \overline{C^{(k)}} \beta^{(k)} \tag{59}$$

with $\overline{C^{(k)}} := \operatorname{Cov}\left(\mathbf{Z}^{(k)}, \mathbf{Z}^{(k)}\right)$. Note that the variance of the estimator has the same expression as in the scalar expectation case (cf. Eq. 19), just replacing $C^{(k)}$ with $\overline{C^{(k)}}$.

Interpretation as averaged covariance matrices The relation $C^{(k)} = R^{(k)} C P^{(k)}$ that defines the $C^{(k)}$ as submatrices of the inter-level covariance matrix is still valid:

$$\overline{C^{(k)}} = R^{(k)} \overline{C} P^{(k)} \tag{60}$$

$$\text{with } \overline{C} := \operatorname{Cov}\left(\mathbf{Z}, \mathbf{Z}\right) \tag{61}$$

The matrix $\overline{C}$ is just the average of the inter-level covariance matrices that would be estimated for each element of the random vectors.

$$\overline{C} = \sum_{i=1}^{n} C^{(k,i)} \tag{62}$$

$$\text{where } \ C^{(k,i)} := \operatorname{Cov}((\mathbf{Z}_{:,i}), (\mathbf{Z}_{:,i})) \tag{63}$$

and where $\mathbf{Z}_{:,i}$ is the i-th column of $\mathbf{Z}$.

Optimal weights and MOSAP From previous paragraphs, it is clear that all results of the scalar expectation case now apply, just replacing the inter-level covariances with averaged inter-level covariances.

For instance, Eq.(36) from the scalar case becomes

$$\beta^{(k)} = m^{(k)} \left(\overline{C^{(k)}}\right)^{-1} R^{(k)} \phi^{-1} \alpha, \tag{64}$$

$$\text{with } \phi = \sum_{k=1}^{K} m^{(k)} P^{(k)} \left(\overline{C^{(k)}}\right)^{-1} R^{(k)} \tag{65}$$

and the MOSAP can be solved by updating $\phi(m)$ in Eq. (39).

Application in large dimensions This approach is tractable for large dimension systems. The most expensive steps have a computational cost that is linear in n:

- For each of the n grid points (or vector elements more generally), estimate an $L \times L$ covariance matrix. Then, take the average of these n matrices. This averaging step should help reducing the sampling noise in the estimation of the covariance matrices.
- As in the scalar expectation case, inverting a few matrices of size at most $L \times L$.

4.3 Field weights

A more refined approximation consists in allowing for different β weights depending on the element consider. This has been introduced by Croci et al. (2023) as a "multi-output" MLBLUE. We don't propose anything new in this section compared to their work.

When the random vector to be estimated can be considered as a discretized random field, the weights β of the multi-output MLBLUE are varying in space. Here, we call this estimator the β-field multilevel estimator:

$$\widehat{\boldsymbol{\mu}}_{\alpha}^{\mathrm{fw}} = \sum_{k=1}^{K} \sum_{\ell \in S^{(k)}} \mathrm{Diag}\left(\boldsymbol{\beta}_{\ell}^{(k)}\right) \widehat{E}^{(k)}\left[\mathbf{Z}_{\ell}\right] \tag{66}$$

where the $\boldsymbol{\beta}_{\ell}^{(k)}$ are now vectors of $\mathbb{R}^n$ and $\mathrm{Diag}\left(\boldsymbol{\beta}_{\ell}^{(k)}\right)$ is the diagonal matrix with diagonal $\boldsymbol{\beta}_{\ell}^{(k)}$.

This class of multilevel estimators includes the previous one (scalar weights). This implies that the optimal β-field estimator is at least as good as the estimator with scalar weights. Nonetheless, it is still a specific class of estimators, that has no reason to include the BLUE.

The minimization problem in this case consists in minimizing a sum of independent problems similar to (22). As a consequence, for a given $i \in \{1, \dots, n\}$, the optimal weights $\beta_{\ell,i}^{(k)}$ are the MLBLUE weights for the scalar random variables $(Z_{\ell,i})_{\ell=1}^{L}$.

Mean value of space-dependent weights Note that the mean values of the $\beta_{\ell}^{(k)}$ vectors differ from the optimal scalar weights (64), due to the non-linearity introduced by the inverse in (64).

Variance of the estimator The variance of the β-field ML estimator is given by

$$\mathbb{V}\left(\widehat{\boldsymbol{\mu}}_{\alpha}^{\mathrm{fw}}(m)\right) = \sum_{i=1}^{n} \alpha^{\intercal} \phi_i(m)^{-1} \alpha \tag{67}$$

$$\text{where } \phi_i(m) := \sum_{k=1}^{K} m^{(k)} P^{(k)} \left(C^{(k,i)}\right)^{-1} R^{(k)}. \tag{68}$$

MOSAP Croci et al. (2023) proposes a solution that minimizes the maximum variance (their equation 21). They also propose variants to minimize, for instance, the total variance:

$$\min_{m \geq 0,\, \mathbf{t} \in \mathbb{R}^n} \|\mathbf{t}\|_1 \quad \text{s.t.} \quad \begin{cases} \begin{pmatrix} \phi_i(m) & \alpha \\ \alpha^{\intercal} & t_i \end{pmatrix} \text{ is positive semi-definite }, \forall i = 1, \dots, n \\ m^{\intercal} c \leq b \\ m^{\intercal} h \geq 1 \end{cases} \tag{69}$$

Application in large dimensions Finding the optimal field weights does not require much more computations than the scalar weight approach. Both require the estimation of the $C^{(k,i)}$ matrices. The β-field approach requires to store $L(L+1)/2$ vectors of size n to store the local covariance matrices. More importantly, it also requires pn inversions of K matrices of size less than $\max_k p^{(k)}$ and of one matrix of size L.

4.4 Field weights with change of basis

Rationale Another still larger (but still suboptimal) class of ML estimator can be defined by applying the estimator in a possibly different space. This approach is motivated by the intuition that a better variance reduction could be obtained with

the field-weight estimator if the data could be linearly transformed into a space where the elements are distributed according to the strength of their interlevel coupling.

For instance, if the low-fidelity models originate from coarse grid simulations, a scale decomposition could provide such a transform, with loose coupling on fine scales and strong coupling on large scales. In this case, there would be a set of $\beta_\ell^{(k)}$ weights for each wave number instead of each vector element.

This class of estimator is especially interesting for two reasons:

1. It is the largest class of estimators that are still computationally tractable in high dimension, i.e. with a computational cost in $\mathcal{O}(n\log(n))$ with respect to the vector size n.

2. It can be interpreted as an optimal post-smoothing, similar to what is done in multigrid methods.

The derivations here may be a bit cumbersome though, so the impatient reader is invited to go directly to section 4.5 on the general multivariate MLBLUE.

Definition Let $\mathbf{W} \in \mathbb{R}^{n\times n}$ be an orthonormal matrix, so that $\mathbf{W}^\intercal\mathbf{W} = \mathbf{W}\mathbf{W}^\intercal = \mathbf{I}_n$. We introduce the class of W-field multilevel estimators:

$$\widehat{\boldsymbol{\mu}}_\alpha^{\mathrm{W}} := \sum_{k=1}^{K} \sum_{\ell\in S^{(k)}} \mathbf{W}\,\mathrm{Diag}\Big(\boldsymbol{\beta}_\ell^{(k)}\Big)\mathbf{W}^\intercal \widehat{E}^{(k)}\big[\mathbf{Z}_\ell\big] \tag{70}$$

$$= \mathbf{W}\sum_{k=1}^{K} \sum_{\ell\in S^{(k)}} \mathrm{Diag}\big(\boldsymbol{\beta}_\ell^{(k)}\big)\,\widehat{E}^{(k)}\left[\mathbf{W}^\intercal\mathbf{Z}_\ell\right]. \tag{71}$$

Relation to the β-field estimators *Intuitively, the optimal W-field estimator should be obtained by applying the optimal β-field estimator to the transformed samples $\mathbf{W}^\intercal\mathbf{Z}_\ell$, and transforming the result back to the physical space. This intuitive result is now properly derived.*

To simplify the notations, let us denote by $\widehat{\boldsymbol{\mu}}^{\mathrm{fw}}(\mathbf{Z},\boldsymbol{\beta})$ a β-field estimator based on samples from $\mathbf{Z}$ and on (possibly non-optimal) field weights $\boldsymbol{\beta}$. $\widehat{\boldsymbol{\mu}}^{\mathrm{fw}}(\mathbf{Z},\boldsymbol{\beta})$ is a possibly biased and possibly sub-optimal estimator for $\boldsymbol{\mu}_\alpha$. Similarly, we denote by $\widehat{\boldsymbol{\mu}}^{\mathrm{W}}(\mathbf{Z},\boldsymbol{\beta},\mathbf{W})$ the W-field estimator based on samples $\mathbf{Z}$ and using field weights $\boldsymbol{\beta}$.

Then equation (71) can be rewritten as

$$\widehat{\boldsymbol{\mu}}^{\mathrm{W}}(\mathbf{Z},\boldsymbol{\beta},\mathbf{W}) := \mathbf{W}\widehat{\boldsymbol{\mu}}^{\mathrm{fw}}(\mathbf{Z}\mathbf{W},\boldsymbol{\beta}) \tag{72}$$

Unbiasedness We first show that unbiased W-field estimators are necessarily associated to unbiased β-field estimator. Let $(P1)$ be the proposition "$\widehat{\boldsymbol{\mu}}^{\mathrm{W}}(\mathbf{Z}, \boldsymbol{\beta}, \mathbf{W})$ is an unbiased estimator of $\mathbb{E}[\mathbf{Z}]^\intercal \alpha$".

$$
\begin{aligned}
(P1) \iff & \quad \mathbb{E}\left[\widehat{\boldsymbol{\mu}}^{\mathrm{W}}(\mathbf{Z}, \boldsymbol{\beta}, \mathbf{W})\right] = \mathbb{E}[\mathbf{Z}]^\intercal \alpha && (73)\\
\iff & \quad \mathbb{E}\left[\mathbf{W}\widehat{\boldsymbol{\mu}}^{\mathrm{fw}}(\mathbf{ZW}, \boldsymbol{\beta})\right] = \mathbb{E}[\mathbf{Z}]^\intercal \alpha && (74)\\
\iff & \quad \mathbf{W}\,\mathbb{E}\left[\widehat{\boldsymbol{\mu}}^{\mathrm{fw}}(\mathbf{ZW}, \boldsymbol{\beta})\right] = \mathbb{E}[\mathbf{Z}]^\intercal \alpha && (75)\\
\iff & \quad \mathbb{E}\left[\widehat{\boldsymbol{\mu}}^{\mathrm{fw}}(\mathbf{ZW}, \boldsymbol{\beta})\right] = \mathbf{W}^\intercal\,\mathbb{E}[\mathbf{Z}]^\intercal \alpha \quad \text{using } \mathbf{W}^\intercal\mathbf{W} = \mathbf{W}\mathbf{W}^\intercal = \mathbf{I}_n && (76)\\
\iff & \quad \mathbb{E}\left[\widehat{\boldsymbol{\mu}}^{\mathrm{fw}}(\mathbf{ZW}, \boldsymbol{\beta})\right] = \mathbb{E}[\mathbf{ZW}]^\intercal \alpha && (77)\\
\iff & \quad (P2) && (78)
\end{aligned}
$$

with $(P2)$: "$\widehat{\boldsymbol{\mu}}^{\mathrm{fw}}(\mathbf{ZW}, \boldsymbol{\beta})$ is an unbiased estimator of $\mathbb{E}\big[\mathbf{ZW}\big]^\intercal \alpha$".

Minimal variance We then show that the mean square errors of two associated estimators are equal.

$$
\begin{aligned}
\mathrm{MSE}\big(\widehat{\boldsymbol{\mu}}^{\mathrm{W}}(\mathbf{Z}, \boldsymbol{\beta}, \mathbf{W}), \mathbb{E}[\mathbf{Z}]^\intercal \alpha\big) &= \mathbb{E}\Big[\big\|\widehat{\boldsymbol{\mu}}^{\mathrm{W}}(\mathbf{Z}, \boldsymbol{\beta}, \mathbf{W}) - \mathbb{E}[\mathbf{Z}]^\intercal \alpha\big\|^2\Big] && (79)\\
&= \mathbb{E}\Big[\big\|\mathbf{W}\widehat{\boldsymbol{\mu}}^{\mathrm{fw}}(\mathbf{ZW}, \boldsymbol{\beta}) - \mathbb{E}[\mathbf{Z}]^\intercal \alpha\big\|^2\Big] && (80)\\
&= \mathbb{E}\Big[\big\|\mathbf{W}^\intercal\mathbf{W}\widehat{\boldsymbol{\mu}}^{\mathrm{fw}}(\mathbf{ZW}, \boldsymbol{\beta}) - \mathbf{W}^\intercal\,\mathbb{E}[\mathbf{Z}]^\intercal \alpha\big\|^2\Big] && (81)\\
&= \mathbb{E}\Big[\big\|\widehat{\boldsymbol{\mu}}^{\mathrm{fw}}(\mathbf{ZW}, \boldsymbol{\beta}) - \mathbb{E}[\mathbf{ZW}]^\intercal \alpha\big\|^2\Big] && (82)\\
&= \mathrm{MSE}\big(\widehat{\boldsymbol{\mu}}^{\mathrm{fw}}(\mathbf{ZW}, \boldsymbol{\beta}), \mathbb{E}[\mathbf{ZW}]^\intercal \alpha\big) && (83)
\end{aligned}
$$

where we used $\mathbf{W}\mathbf{W}^\intercal = \mathbf{I}_n$ to obtain (81) and $\mathbf{W}^\intercal\mathbf{W} = \mathbf{I}_n$ to obtain (82).

Since the unbiased estimators have equal mean square errors, the best unbiased W-field estimator is associated to the best unbiased β-field estimator through (72). All results valid for the β-field estimator (optimal values, sample allocation) can thus be applied here, by replacing random vectors $\mathbf{Z}_\ell$ by $\mathbf{W}^\intercal\mathbf{Z}_\ell$.

Numerical cost of applying the estimator To limit the number of transforms, the W-field ML estimator can be applied as

$$
\widehat{\boldsymbol{\mu}}_\alpha^{\mathrm{W}} = \mathbf{W}\sum_{k=1}^{K}\sum_{\ell\in S^{(k)}} \mathrm{Diag}\left(\boldsymbol{\beta}_\ell^{(k)}\right)\mathbf{W}^\intercal\widehat{E}^{(k)}\big[\mathbf{Z}_\ell\big] \tag{84}
$$

which counts p forward transforms and one inverse transform (in the MLMC case, $2L$ transforms. See page 7 for the definition of p).

Numerical cost of estimating the optimal weights The estimation of the optimal weights requires the computation of element-wise covariance matrices $C^{(k,i)}$. The simplest approach consists in estimating the $C^{(k,i)}$ from a set of $N \times L$ transformed and coupled realizations, with N a large sampling size. The transformation of this training ensemble into the W-space requires $N \times L$ forward transforms.

Choice of W This estimator depends on the choice of $\mathbf{W}$. With $\mathbf{W} = \mathbf{I}_n$, we retrieve the β-field estimator. With a Fourier basis, we get the optimal spectral filters.

Note that if $\mathbf{W}$ is allowed to vary, the W-field estimators encompass all estimators with real valued symmetric matrices that are simultaneously diagonalizable. This is less general than the BLUE introduced in section 4.5, where the matrix weights are generally not symmetric.

4.5 Matrix weights – the multidimensional MLBLUE

The general MLBLUE for a random vector estimation has matrix weights, as mentioned in Croci et al. (2023). We follow the same approach as SU20 for the derivation of the MLBLUE in this context. A variance minimization approach is also feasible and would yield the same results (see appendix A).

Notations Some notations need to be updated for this section:

- The random vectors $\mathbf{Z}_1, \dots, \mathbf{Z}_L$ on each levels can now have different sizes n_ℓ possibly all different from n. For instance, if the low-fidelity samples originate from simulations on coarse grids, there is no requirement to interpolate them back to the grid of the fine level L. To shorten notations, we define:

$$N := \sum_{\ell=1}^{L} n_\ell \tag{85}$$

$$n^{(k)} := \sum_{\ell \in S(k)} n_\ell \tag{86}$$

- $\mathbf{Z}$ is now the concatenation of the L random vectors $\mathbf{Z}_\ell \in \mathbb{R}^{n_\ell}$.

$$\mathbf{Z} := \left(\mathbf{Z}_1^\intercal \cdots \mathbf{Z}_L^\intercal\right)^\intercal \in \mathbb{R}^N \tag{87}$$

$$\boldsymbol{\mu} := \mathbb{E}\big[\mathbf{Z}\big] \tag{88}$$

- The selection and extension operators are extended accordingly, to select only the part $\mathbf{Z}^{(k)}$ of $\mathbf{Z}$ which is relevant to some coupling group k.

$$\mathbf{Z}^{(k)} := \left(\cdots \mathbf{Z}_\ell^\intercal \cdots\right)_{\ell \in S^{(k)}}^\intercal \in \mathbb{R}^{n^{(k)}} \tag{89}$$

$$\mathbf{R}^{(k)} \text{ of size } n^{(k)} \times N \text{ such that } \mathbf{R}^{(k)}\mathbf{Z} = \mathbf{Z}^{(k)} \tag{90}$$

$$\mathbf{P}^{(k)} := \mathbf{R}^{(k)\intercal} \tag{91}$$

$$\tag{92}$$

- The α coefficients are extended to a matrix $\boldsymbol{\alpha}$ representing a linear map from $\mathbf{R}^N$ to $\mathbf{R}^n$. In the most common case, we are interested in the high-fidelity level and $n = n_L$. In this case, $\boldsymbol{\alpha}$ is the selection operator for this level.

$$\boldsymbol{\alpha} := \left(\mathbf{0}_{n\times n_1} \ldots \mathbf{0}_{n\times n_{L-1}} \mathbf{I}_n\right) \tag{93}$$

$$\tag{94}$$

  The quantity of interest here is $\boldsymbol{\mu}_\alpha := \boldsymbol{\alpha}\boldsymbol{\mu} \in \mathbb{R}^n$.

Normal equations All samples are concatenated in a big column vector of "observations" $\underline{\mathbf{Z}} := \left(\left(\mathbf{Z}^{(k),i}\right)_{i=1}^{m^{(k)}}\right)_{k=1}^K$, where $\mathbf{Z}^{(k),i} \in \mathbb{R}^{n^{(k)}}$ is the i-th (random) sample on coupling group k. The size of $\underline{\mathbf{Z}}$ is $\sum_{k=1}^K m^{(k)} n^{(k)}$.

The "observation operator" relating $\boldsymbol{\mu}$ and $\underline{\mathbf{Z}}$ is the column block vector $\mathbf{H} := \left(\left(\mathbf{R}^{(k)}\right)_{i=1}^{m^{(k)}}\right)_{k=1}^K$. Then

$$\underline{\mathbf{Z}} = \mathbf{H}\boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{95}$$

$$\text{with} \quad \boldsymbol{\epsilon} := \underline{\mathbf{Z}} - \mathbf{H}\boldsymbol{\mu} \tag{96}$$

We have the following properties about the noise $\boldsymbol{\epsilon}$.

$$\mathbb{E}[\boldsymbol{\epsilon}] = 0 \tag{97}$$

$$\mathrm{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) = \mathrm{Cov}\left(\underline{\mathbf{Z}}, \underline{\mathbf{Z}}\right) \tag{98}$$

$$= \mathrm{Diag}_{k=1}^K\left(\mathrm{Diag}_{i=1}^{m^{(k)}}\left(\mathbf{C}^{(k)}\right)\right) \tag{99}$$

$$\text{with} \quad \mathbf{C}^{(k)} := \mathrm{Cov}\left(\mathbf{Z}^{(k)}, \mathbf{Z}^{(k)}\right) \tag{100}$$

Invertibility of the covariance matrices Thereafter, we assume that the matrices $\mathbf{C}^{(k)}$ are non-singular.

This may not be the case in some cases, and some care should be taken in defining the low-fidelity samples. For instance, the assumption is not valid if some

low fidelity samples on level ℓ are coarse grid simulations linearly interpolated to a finer grid. In this situation, the extrapolated elements in $\mathbf{Z}_\ell$ can be expressed as a linear combination of the other elements it was extrapolated from, so that $\mathbf{C}^{(k)}$ has a non zero kernel and is singular.

To meet the non-singularity assumption in this case, the low fidelity simulations on the coarse grid should be used as is, without any interpolation. It is the role of the MLBLUE to compute the optimal linear transform that best maps samples from this low-fidelity level to the finest grid. The matrix weights here act not only as optimal multilevel weights, but also as interpolators and smoothers.

Associated generalized least-squares problem

$$\min_{\boldsymbol{\mu}\in\mathbf{R}^N} \|\underline{\mathbf{Z}} - \mathbf{H}\boldsymbol{\mu}\|^2_{\mathrm{Cov}(\boldsymbol{\epsilon},\boldsymbol{\epsilon})^{-1}} \tag{101}$$

which could be decomposed as a sum over the coupling group k, as a consequence of the block-diagonal structure of $\mathrm{Cov}(\boldsymbol{\epsilon},\boldsymbol{\epsilon})^{-1}$.

BLUE for the expectation of a random vector The solution is given by

$$\widehat{\boldsymbol{\mu}}^{\mathrm{mat},\star} = \left(\mathbf{H}^\intercal \mathrm{Cov}(\boldsymbol{\epsilon},\boldsymbol{\epsilon})^{-1}\mathbf{H}\right)^{-1} \mathbf{H}^\intercal \mathrm{Cov}(\boldsymbol{\epsilon},\boldsymbol{\epsilon})^{-1}\underline{\mathbf{Z}} \tag{102}$$

$$= \boldsymbol{\phi}^{-1}\mathbf{y} \tag{103}$$

$$\text{with} \quad \boldsymbol{\phi} := \sum_{k=1}^{K} m^{(k)}\mathbf{P}^{(k)} \left(\mathbf{C}^{(k)}\right)^{-1} \mathbf{R}^{(k)} \tag{104}$$

$$\text{and} \quad \mathbf{y} := \sum_{k=1}^{K} m^{(k)}\mathbf{P}^{(k)} \left(\mathbf{C}^{(k)}\right)^{-1} \widehat{E}^{(k)}\left[\mathbf{Z}^{(k)}\right] \tag{105}$$

Partial estimation The BLUE for $\boldsymbol{\mu}_\alpha$ is actually $\boldsymbol{\alpha}\widehat{\boldsymbol{\mu}}^{\mathrm{mat},\star}$.

Matrix weights The previous equations can be expanded to evidence the multilevel structure of the estimator.

$$\boldsymbol{\alpha}\widehat{\boldsymbol{\mu}}^{\mathrm{mat},\star} = \sum_{k=1}^{K} \boldsymbol{\beta}^{(k)} \widehat{E}^{(k)}\left[\mathbf{Z}^{(k)}\right] \tag{106}$$

$$= \sum_{k=1}^{K} \sum_{\ell\in S^{(k)}} \boldsymbol{\beta}_\ell^{(k)} \widehat{E}^{(k)}\left[\mathbf{Z}_\ell\right] \tag{107}$$

$$\boldsymbol{\beta}^{(k)} = m^{(k)}\boldsymbol{\alpha}\boldsymbol{\phi}^{-1}\mathbf{P}^{(k)} \left(\mathbf{C}^{(k)}\right)^{-1} \in \mathbb{R}^{n\times n^{(k)}} \tag{108}$$

$$= \left(\cdots\ \boldsymbol{\beta}_\ell^{(k)} \cdots\right)_{\ell\in S^{(k)}} \tag{109}$$

Variance and sample allocation The approach of Croci et al. (2023) to solve the MOSAP via semdidefinite programming is not directly applicable to this case. Some work on extending it would be of interest.

For a given selection of groups of levels, the variance of the ML matrix-weighted estimator with optimal matrix weights is given by

$$\operatorname{Tr}\operatorname{Cov}\left(\widehat{\boldsymbol{\mu}}^{\mathrm{mat},\star}, \widehat{\boldsymbol{\mu}}^{\mathrm{mat},\star}\right) = \operatorname{Tr}\left(\boldsymbol{\alpha}\boldsymbol{\phi}^{-1}(m)\boldsymbol{\alpha}^{\intercal}\right) \tag{110}$$

which can be used to find the optimal sample allocation m for this specific choice of levels and coupling structure.

With suboptimal matrix weights $\boldsymbol{\beta}^{(k)}$, the variance is given by

$$\operatorname{Tr}\operatorname{Cov}\left(\widehat{\boldsymbol{\mu}}^{\mathrm{mat}}, \widehat{\boldsymbol{\mu}}^{\mathrm{mat}}\right) = \sum_{k=1}^{K} m^{(k)} \operatorname{Tr}\left(\boldsymbol{\beta}^{(k)}\mathbf{C}^{(k)}\left(\boldsymbol{\beta}^{(k)}\right)^{\intercal}\right) \tag{111}$$

Remark In the case of a weighted MLMC, the no-bias condition implies that last matrix weight $\boldsymbol{\beta}_{L,K}$ is $\mathbf{I}_n$.

Computational cost This approach is untractable as is for large-dimension systems. Indeed, estimating the matrix weights requires:

- Estimating the covariance matrix $\mathbf{C} := \operatorname{Cov}(\mathbf{Z}, \mathbf{Z})$ of size N by N, which is a $\mathcal{O}\left(n^2\right)$ task. The covariance matrices $\mathbf{C}^{(k)}$ can then be extracted as $\mathbf{C}^{(k)} = \mathbf{R}^{(k)}\mathbf{C}\mathbf{P}^{(k)}$. Some of these $\mathbf{C}^{(k)}$ are invertible only if $\mathbf{C}$ is estimated with at least $\max_k n^{(k)} + 1$ samples, which may be very expensive. A possible workaround to the singularity of $\mathbf{C}$ could be using some regularization technique such as covariance localization or ridging.
- Inverting the $\boldsymbol{\phi}$ matrix, of size N, and each $\mathbf{C}^{(k)}$ matrix of size $n^{(k)}$, or solving linear systems of associated sizes if the weights are directly applied to some MC estimate.

Making it tractable in large dimensions There are various ways forward to simplify this multilevel estimator to make it tractable in large dimension. If the simulations are attached to a physical space with some notion of distance, the interpretation of the matrix weights as interpolators and smoothers suggest that the $\boldsymbol{\beta}_{\ell}^{(k)}$ matrices could be imposed as sparse. This means that the estimation of an element on the fine grid should only involve low-fidelity elements that are within some maximum distance. The matrix weights could be further simplified by imposing some invariance by translation, or have some periodicity based on the underlying grids.

Alternatively, the simplification could be done the other way round. The inter-level and inter-element covariance matrices $\mathbf{C}^{(k)}$ could be computed only on pair of points within a given distance. Beyond this distance, the covariances would be assumed to be zero. This could be an interesting avenue for future research.

Relation with other multilevel estimators The multilevel estimators introduced in sections 4.2 to 4.4 can be considered as special cases of the general matrix-weight estimator.

- The estimator with scalar weights is restricting itself to $\boldsymbol{\beta}_\ell^{(k)}$ matrices of the form $\beta_\ell^{(k)}\mathbf{I}_n$.
- The β-field estimator is restricting itself to $\boldsymbol{\beta}_\ell^{(k)}$ matrices of the form $\mathrm{Diag}\left(\boldsymbol{\beta}_\ell^{(k)}\right)$ (where $\boldsymbol{\beta}_\ell^{(k)}$ is now a vector).
- The W-field estimator is restricting itself $\boldsymbol{\beta}_\ell^{(k)}$ matrices that are all diagonalizable in the basis $\mathbf{W}$.
- The tractable approaches mentioned in previous paragraph would restrict to sparse matrix weights, or kernel-based matrices.

5 Estimation of a scalar covariance

Let $X = (X_\ell)_{\ell=1}^L$ and $Y = (Y_\ell)_{\ell=1}^L$ be random vectors gathering the scalar random variables $X_1, \dots, X_L$ and $Y_1, \dots, Y_L$. We group the covariances of X_ℓ and Y_ℓ for each ℓ in the vector $c := (\mathrm{Cov}(X_\ell, Y_\ell))_{\ell=1}^L$. We are interested in a linear combination of the elements of c, of the kind $\alpha^\intercal c$, with $\alpha \in \mathbb{R}^n \setminus \{0\}$. In practice, $\alpha = (0 \dots 0\ 1) \in \mathbb{R}^L$.

Given some some coupling structure $\left(S^{(k)}, m^{(k)}\right)_{k=1}^K$, we are looking for the best unbiased estimator of $\alpha^\intercal c$ of the form

$$\widehat{C}_\alpha^{\mathrm{ML}} = \sum_{k=1}^K \sum_{\ell \in S^{(k)}} \beta_\ell^{(k)} \widehat{C}^{(k)}(X_\ell, Y_\ell) \tag{112}$$

where $\widehat{C}^{(k)}(X_\ell, Y_\ell)$ is the sample covariance estimator of X_ℓ and Y_ℓ based on $m^{(k)}$ coupled samples.

$$\widehat{C}^{(k)}(X_\ell, Y_\ell) := \frac{m^{(k)}}{m^{(k)} - 1} \widehat{E}^{(k)}\left(\left(X_\ell - \widehat{E}^{(k)}(X_\ell)\right)\left(Y_\ell - \widehat{E}^{(k)}(Y_\ell)\right)\right) \tag{113}$$

This is an unbiased estimator for $\mathrm{Cov}(X_\ell, Y_\ell)$.

A linear estimator? The multilevel MC estimators of the form (112) depend quadratically on the samples, and as such are not truly linear estimators. A multilevel best *linear* unbiased estimator for the covariance does not exist in general. However, estimators (112) are still linear in the MC estimators they combine, which is why we use the name MLBLUE for covariance estimators as well. The regression approach could be used here to derive the MLBLUE, but the MC estimators on each coupling groups should be used instead of the samples.

Relation to the expectation case This problem is very similar to the problem faced in section 3 for the estimation of the expectation. The two important points needed to extend the results are the unbiasedness of the MC estimators involved, and an expression of their covariances.

- $\mathrm{Diag}\left(\widehat{C}^{(k)}\left(X^{(k)}, Y^{(k)}\right)\right)$ is an unbiased estimator for $R^{(k)}c$, similar to the expectation case where $\widehat{E}^{(k)}\left[Z^{(k)}\right]$ is an unbiased estimator of $R^{(k)}\mu$.
- However, the covariance of MC estimators for the covariances does not simplify as well as the MC estimators for the mean.

Let $\mathbb{C}^{(k)}$, of size $p^{(k)}$ by $p^{(k)}$, denote the covariance of the MC estimator on group k with $m^{(k)}$ samples. It differs from the covariance $C^{(k)}$ of the random vectors $Z^{(k)}$. For the estimation of the mean, we have

$$\mathbb{C}^{(k)} = \mathrm{Cov}\left(\widehat{E}^{(k)}\left[Z^{(k)}\right], \widehat{E}^{(k)}\left[Z^{(k)}\right]\right) \tag{114}$$

$$= 1/m^{(k)} C^{(k)} \tag{115}$$

For the estimation of the covariance, we have

$$\mathbb{C}^{(k)} := \mathrm{Cov}\left(\mathrm{Diag}\left(\widehat{C}^{(k)}\left(X^{(k)}, Y^{(k)}\right)\right), \mathrm{Diag}\left(\widehat{C}^{(k)}\left(X^{(k)}, Y^{(k)}\right)\right)\right) \tag{116}$$

which is expanded later in this section.

MLBLUE results expressed as a function of the $\mathbb{C}^{(k)}$ The solution to the constrained minimization problem is given by

$$\beta^{(k)} = \left(\mathbb{C}^{(k)}\right)^{-1} R^{(k)} \phi^{-1} \alpha \tag{117}$$

$$\text{with } \phi := \sum_{k=1}^{K} P^{(k)} \left(\mathbb{C}^{(k)}\right)^{-1} R^{(k)}. \tag{118}$$

The variance of the MLBLUE is $\alpha^\intercal \phi^{-1} \alpha$. Note that this paragraph is valid for the estimation of the expectation, for the estimation of the covariance, but also for the estimation of any scalar statistic which admits unbiased Monte Carlo estimators. @Paul,Selime: Should this be underlined by transforming the section into *Estimation of a scalar statistics*?

Estimating the $\mathbb{C}^{(k)}$ Covariances $\mathbb{C}^{(k)}$ can be expressed as a function of the first centered-moments of $R^{(k)}X$ and $R^{(k)}Y$, isolating the dependence on the sample size $m^{(k)}$. Let's drop the k for the sake of clarity, and focus on the covariance matrix $\mathbb{C}$ associated to X and Y. It can be shown (equation 9 of M15a) that for $\ell, \ell' \in \{1, \ldots, L\}$, element ℓ, ℓ' of $\mathbb{C}$ is

$$\begin{aligned}\mathbb{C}_{\ell,\ell'} = &\frac{\mathbb{M}^4\left[X_\ell, X'_\ell, Y_\ell, Y'_\ell\right]}{m^{(k)}} - \frac{\mathrm{Cov}\left(X_\ell, Y_\ell\right)\mathrm{Cov}\left(X_{\ell'}, Y_{\ell'}\right)}{m^{(k)}} \\ &+ \frac{\mathrm{Cov}\left(X_\ell, Y_{\ell'}\right)\mathrm{Cov}\left(Y_\ell, X_{\ell'}\right) + \mathrm{Cov}\left(X_\ell, X_{\ell'}\right)\mathrm{Cov}\left(Y_\ell, Y_{\ell'}\right)}{m^{(k)}(m^{(k)}-1)}\end{aligned} \tag{119}$$

with $\mathbb{M}^4\left[X_1, X_2, X_3, X_4\right] := \mathbb{E}\left[\Pi_{i=1}^4 \left(X_i - \mathbb{E}\left[X_i\right]\right)\right]$.

In the case of variance estimation, when $X = Y$, equation (119) simplifies to

$$\mathbb{C}_{\ell,\ell'} = \frac{\mathbb{M}^4\left[X_\ell, X'_\ell\right]}{m^{(k)}} - \frac{\mathbb{V}\left(X_\ell\right)\mathbb{V}\left(X_{\ell'}\right)}{m^{(k)}} + \frac{2\,\mathrm{Cov}\left(X_\ell, X_{\ell'}\right)^2}{m^{(k)}(m^{(k)}-1)} \tag{120}$$

with $\mathbb{M}^4[X_1, X_2] := \mathbb{M}^4[X_1, X_1, X_2, X_2]$ (the definition of $\mathbb{M}^4$ depends on the number of parameters it is given).

MOSAP The semidefinite programming approach to the MOSAP does not extends naturally here, as the matrix $\phi(m)$ no longer depends linearly on the sample sizes m. Note this will be the case for all multilevel estimators of the covariance introduced in this note, including multilevel estimators of covariance matrices.

This could be circumvented by minimizing an upper bound of the variance that linearly depends on the inverses of the samples sizes, as done ny Mycek and De Lozzo (2019).

Another solution that does not introduce approximations is given by falling back to the constrained minimization of the non linear variance $\alpha^\intercal \phi(m)^{-1}\alpha$, or by solving the non-linear semidefinite problem of Croci et al. (2023).

Computational cost

- In practice, all the involved matrices are of size at most L (*i.e.* 2 or 3).
- Ensemble estimates of fourth-order moments are more noisy than ensemble covariances. As a consequence, the ensemble sizes used to estimate $\mathbb{C}$ should be larger here than in section 3 to reach a similar robustness.
- $\mathbb{C}$ is a symmetric matrix, so only $L(L+1)/2$ elements really need to be estimated. This number can be even more reduced by noting that covariances between levels that are not coupled (ℓ, ℓ' so that $\forall k,\ \ell \in S^{(k)} \Rightarrow \ell' \notin S^{(k)}$) need not be computed.

6 Estimation of a covariance matrix

6.1 The problem

How do the results of previous section extend to the multidimensional case? The general MLBLUE for the estimation of a covariance matrix involves fourth-order tensors linearly combining the entries of several covariance matrix estimators into one matrix. We don't derive it here, as it is computationnally prohibitively expensive in high-dimension applications. Instead, we focus on multilevel estimators of the covariance matrix that are both simpler and computationally affordable.

The $\mathbf{X}_\ell$ are now random vectors of $\Omega \to \mathbb{R}^n$. Note that they are imposed to have the same dimension, which would not be the case with the general MLBLUE. They are concatenated in the random vector $\mathbf{X} := (\mathbf{X}_1^\intercal, \dots, \mathbf{X}_L^\intercal)^\intercal$. We are interested in the covariance matrix $\mathrm{Cov}(\mathbf{X}, \mathbf{X})$ of size nL.

Partial estimation More specifically, we are often interested in estimating only some blocks of $\mathrm{Cov}(\mathbf{X}, \mathbf{X})$, for instance the last block $\mathrm{Cov}(\mathbf{X}_L, \mathbf{X}_L)$. To extend the formalism to this kind of situation, we focus on estimating $\mathbf{C}^\alpha := \sum_{\ell=1}^L \alpha_\ell \mathrm{Cov}(\mathbf{X}_\ell, \mathbf{X}_\ell)$ for some non-zero vector $\alpha \in \mathbb{R}^L$.

Generalized variance for a matrix estimator We choose the Frobenius norm to measure mean square errors for matrix estimators, as in M15a. The choice of the Frobenius norm is associated to the following bias-variance decomposition:

$$\mathrm{MSE}\left(\widehat{\mathbf{B}}, \mathbf{B}\right) = \sum_{i=1}^n \sum_{j=1}^n \mathbb{V}(\widehat{B}_{ij}) + \left\| \mathbb{E}\left[\widehat{\mathbf{B}} - \mathbf{B}\right] \right\|_{\mathrm{F}}^2 \tag{121}$$

Hereafter, we overload the variance operator and define

$$\mathbb{V}(\widehat{\mathbf{B}}) := \sum_{i=1}^n \sum_{j=1}^n \mathbb{V}(\widehat{B}_{ij}) \tag{122}$$

for any matrix-valued random variable $\widehat{\mathbf{B}}$.

6.2 Scalar weights

We first consider the case of scalar weights, one for each MC estimator, common to all matrix elements in the estimator.

Class of estimators

$$\widehat{\mathbf{C}}^{\mathrm{ML}} = \sum_{k=1}^{K} \sum_{\ell \in S^{(k)}} \beta_\ell^{(k)} \widehat{C}^{(k)} \left(\mathbf{X}_\ell, \mathbf{X}_\ell \right) \tag{123}$$

where $\beta_\ell^{(k)}$ is a scalar.

Optimal scalar weights The optimal weights can be retrieved very similarly to what is done for multidimensional expectation in section 4.2.

The same relation (53) guarantees the unbiasedness of the estimator. It can be used as a constraint to minimize the variance of the estimator, given by a relationship similar to equation (59).

$$\mathbb{V}\left(\widehat{\mathbf{C}}^{\mathrm{ML}}\right) = \sum_{k=1}^{K} \left(\beta^{(k)}\right)^{\intercal} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{C}^{(k,ij)} \right) \beta^{(k)} \tag{124}$$

where $\mathbb{C}^{(k,ij)}$ is the covariance matrix of size $p^{(k)} \times p^{(k)}$ for Monte Carlo covariance estimators $\widehat{C}^{(k)}(X_{\ell,i}, X_{\ell,j})$, $\ell \in S^{(k)}$.

The optimal weights are then given by equation (117), replacing $\mathbb{C}^{(k)}$ with $\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{C}^{(k,ij)}$, or equivalently with $1/n^2 \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{C}^{(k,ij)}$.

Estimating the average $\mathbb{C}^{(k,ij)}$ The expression of the average inter-level covariance matrix of MC estimators directly follows from the expression in the scalar case (section 5).

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{C}_{\ell,\ell'}^{(k,ij)} = \sum_{i=1}^{n} \sum_{j=1}^{n} \Bigg(& \frac{\mathbb{M}^4 \left[X_{\ell,i}, X_{\ell',i}, X_{\ell,j}, X_{\ell',j} \right]}{m^{(k)}} \\
& - \frac{\mathrm{Cov}\left(X_{\ell,i}, X_{\ell,j}\right) \mathrm{Cov}\left(X_{\ell',i}, X_{\ell',j}\right)}{m^{(k)}} \\
& + \frac{\mathrm{Cov}\left(X_{\ell,i}, X_{\ell',j}\right) \mathrm{Cov}\left(X_{\ell,j}, X_{\ell',i}\right) + \mathrm{Cov}\left(X_{\ell,i}, X_{\ell',i}\right) \mathrm{Cov}\left(X_{\ell,j}, X_{\ell',j}\right)}{m^{(k)}(m^{(k)} - 1)} \Bigg)
\end{aligned} \tag{125}$$

MOSAP As explained in section 5, the semidefinite programming approach to the MOSAP does not extend naturally here.

Computational cost Estimating $\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{C}^{(k,ij)}$ directly is not tractable in high dimension, as the number of operations is quadratic in the grid size n. Fortunately, it is possible to reduce the cost of this estimation from $\mathcal{O}\left(n_e n^2\right)$ to $\mathcal{O}\left(n_e^2 n\right)$, where n_e is the size of the ensemble used to estimate centered statistics in the preprocessing step (see appendix B).

Sample allocation for MLMC The expression of the variance of a general estimator given here can be used to estimate the optimal sample allocation for the MLMC estimator of a covariance matrix. This is detailed in appendix C.

6.3 Matrix field weights

Class of estimators

$$\widehat{\mathbf{C}}^{\mathrm{ML}} = \sum_{k=1}^{K} \sum_{\ell \in S^{(k)}} \boldsymbol{\beta}_\ell^{(k)} \circ \widehat{C}^{(k)} \left(\mathbf{X}_\ell, \mathbf{X}_\ell \right) \tag{126}$$

where $\boldsymbol{\beta}_\ell^{(k)}$ is an $n \times n$ matrix and $\circ$ denotes the Schur (element-wise) product.

Unbiasedness, Variance, Optimal weights From the definition of the generalized variance derived from the Frobenius norm, the problem can be decomposed as independent estimations of n^2 scalar covariances. See section 5 for more details.

Numerical cost Unaffordable as such, since it requires the estimation of n^2 scalar coefficients.

Relation to covariance localization The use of a Schur product applied to a covariance matrix in equation (126) reminds of covariance localization in data assimilation (e.g. Lorenc, 2003). This regularization technique consists in element-wise multiplication of an ensemble covariance matrix with a parametrized distance-dependent correlation matrix $\mathbf{L}$:

$$\mathbf{L} \circ \widehat{C}^{(k)} \left(\mathbf{X}_\ell, \mathbf{X}_\ell \right) \tag{127}$$

Localization differs from what has been considered so far, has it yields a *biased* estimator of the covariance matrix wherever $\mathbf{L}$ is not 1. Optimizing the variance without the bias constraint makes no sense, but optimizing the MSE does, since it includes a bias and a variance contribution. This is one of the ideas presented in M15b, which we extend to the multilevel case in next section.

6.4 Optimal localization

We propose here an extension of the optimal localization theory of M15a and M15b to multilevel covariance estimators. We first derive the results in the line of the previous sections. An aptempt to propose a derivation based on the approach of M15a and M15b (see also Ménétrier, 2020) is proposed in appendix D. Though both approaches yield the same practical results, the associated assumptions and interpretations are slightly different.

Contrarily to previous sections, we no longer minimize the variance under a no-bias constraint. The localization makes the covariance estimator biased, so we rather minimize the mean squared error of the localized multilevel estimator.

6.4.1 General case

To simplify the notations, we note $\mathbf{B}_\ell := \mathrm{Cov}(\mathbf{X}_\ell, \mathbf{X}_\ell)$ and $\widetilde{\mathbf{B}}_\ell^{(k)} := \widehat{C}^{(k)}(\mathbf{X}_\ell, \mathbf{X}_\ell)$. From the unbiasedness of the Monte Carlo covariance estimator, we have $\mathbb{E}\left[\widetilde{\mathbf{B}}_\ell^{(k)}\right] = \mathbf{B}_\ell$.

Class of estimators

$$\widehat{\mathbf{B}}^{\mathrm{ML}} = \sum_{k=1}^{K} \sum_{\ell \in S^{(k)}} \mathbf{L}_\ell^{(k)} \circ \widetilde{\mathbf{B}}_\ell^{(k)} \tag{128}$$

where $\mathbf{L}_\ell^{(k)}$ is an $n \times n$ matrix, without any imposed structure at this stage.

Minimizing the MSE We want to minimize the mean square error of the localized covariance estimator:

$$\mathrm{MSE}\left(\widehat{\mathbf{B}}^{\mathrm{ML}}, \mathbf{B}_L\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}\left[\left(\sum_{k=1}^{K} \sum_{\ell \in S^{(k)}} L_{\ell,ij}^{(k)} \circ \widetilde{B}_{\ell,ij}^{(k)} - B_{L,ij}\right)^2\right] \tag{129}$$

The function to minimize is the sum of n^2 independent cost functions. We now focus on one independent sub-problem i, j. To ease the notations, we drop the ij indexes and denote $L_\ell^{(k)} := L_{\ell,ij}^{(k)}$, $\widetilde{B}_\ell^{(k)} := \widetilde{B}_{\ell,ij}^{(k)}$ and $B_\ell := B_{\ell,ij}$. The problem reads as

$$\min_{L_\ell^{(k)},\, 1 \le k \le K,\, \ell \in S^{(k)}} \mathbb{E}\left[\left(\sum_{k=1}^{K} \sum_{\ell \in S^{(k)}} L_\ell^{(k)} \widetilde{B}_\ell^{(k)} - B_L\right)^2\right] \tag{130}$$

To further simplify the notations, we define stacked vectors containing the information from all levels in all coupling groups:

$$\underline{L} := \left(\left(L_\ell^{(k)}\right)_{\ell \in S^{(k)}}\right)_{k=1}^{K} \in \mathbb{R}^p \tag{131}$$

$$\underline{\widetilde{B}} := \left(\left(\widetilde{B}_\ell^{(k)}\right)_{\ell \in S^{(k)}}\right)_{k=1}^{K} \in \mathbb{R}^p \tag{132}$$

$$\underline{B} := \left(\left(B_\ell\right)_{\ell \in S^{(k)}}\right)_{k=1}^{K} \in \mathbb{R}^p \tag{133}$$

with $p = \sum_{k=1}^{K} p^{(k)}$.

Equation (130) becomes

$$\begin{aligned}\min_{\underline{L}} \mathbb{E}\left[\left(\underline{L}^\intercal \underline{\widetilde{B}} - B_L\right)^2\right] &= \min_{\underline{L}} \mathbb{E}\left[\left(\underline{L}^\intercal \underline{\widetilde{B}}\right)\left(\underline{\widetilde{B}}^\intercal \underline{L}\right) - 2B_L\left(\underline{\widetilde{B}}^\intercal \underline{L}\right) + B_L^2\right] \\ &= \min_{\underline{L}} \underline{L}^\intercal \mathbb{E}\left[\underline{\widetilde{B}}\underline{\widetilde{B}}^\intercal\right]\underline{L} - 2B_L \mathbb{E}\left[\underline{\widetilde{B}}^\intercal\right]\underline{L} + B_L^2 \\ &= \min_{\underline{L}} \underline{L}^\intercal \mathbb{E}\left[\underline{\widetilde{B}}\underline{\widetilde{B}}^\intercal\right]\underline{L} - 2B_L \underline{B}^\intercal \underline{L} \end{aligned} \tag{134}$$

Setting the gradient with respect to $\underline{L}$ to zero yields the optimality criterion

$$\mathbb{E}\left[\underline{\widetilde{B}}\underline{\widetilde{B}}^\intercal\right]\underline{L} = B_L \underline{B} \tag{135}$$

Invertibility of $\mathbb{E}\left[\underline{\widetilde{B}}\underline{\widetilde{B}}^\intercal\right]$ The uniqueness of the optimal localization is not guaranteed, in particular if the matrix $\mathbb{E}\left[\underline{\widetilde{B}}\underline{\widetilde{B}}^\intercal\right]$ on the left-hand side is not invertible. This should not be a problem though, as cases of non-invertibility are related to flat gradient in the MSE, which means to variations of $\underline{L}$ which don't impact the MSE. In practice, the non-uniqueness of these cases i, j can be decided from the neighbouring pairs of points. For instance, parametric correlation matrices can be fit to these raw optimal localizations, to ensure the positive semidefiniteness of the localization matrices.

Sample or asymptotic quantities The left hand-side matrix could be expressed as a function of the sample size and asymptotic quantities. We could then use a large independent ensemble to estimate these asymptotic quantities. In practice, we don't have such large ensembles. A possible workaround is to express everything in terms of expectations of sample quantities, and to exploit the structure of the localization matrix.

6.4.2 Imposing some structure to the localization matrix

In practice, the localization matrix has some predefined structure, both for ease of estimation and ease of use. The state space $\mathbb{R}^n$ is attached to some physical space, three-dimensional for instance. Assuming for instance horizontal homogeneity and isotropy, we can define an equivalence relation between pairs (i, j), associating pairs that should have the same localization. The space of pairs $\mathbb{R}^n \times \mathbb{R}^n$ is thus partitioned into equivalence classes.

$$L_{\ell,ij}^{(k)} = L_{\ell,\mathcal{C}}^{(k)} \text{ where } \mathcal{C} = [(i,j)] \text{ is the equivalence class of } (i,j) \tag{136}$$

In this context, equation (129) reads as

$$\mathrm{MSE}\left(\widehat{\mathbf{B}}^{\mathrm{ML}}, \mathbf{B}_L\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\left(\sum_{k=1}^{K}\sum_{\ell\in S^{(k)}} L^{(k)}_{\ell,[(i,j)]} \circ \widetilde{B}^{(k)}_{\ell,ij} - B_{L,ij}\right)^2\right] \tag{137}$$

where the localization $L^{(k)}_{\ell,ij}$ has been replaced by $L^{(k)}_{\ell,[(i,j)]}$. This minimization problem is associated to independent problems for each class $\mathcal{C} = [(i,j)]$.

$$\min_{L^{(k)}_{\ell,\mathcal{C}}, 1\leq k\leq K, \ell\in S^{(k)}} \sum_{(i,j)\in\mathcal{C}} \mathbb{E}\left[\left(\sum_{k=1}^{K}\sum_{\ell\in S^{(k)}} L^{(k)}_{\ell,\mathcal{C}}\widetilde{B}^{(k)}_{\ell,ij} - B_{L,ij}\right)^2\right] \tag{138}$$

As previously done, we define stacked vectors of $\mathbb{R}^p$: $\underline{L}$, $\underline{\widetilde{B}_{ij}}$ and $\underline{B_{ij}}$ for $(i,j)\in\mathcal{C}$. Expanding the MSE gives

$$\begin{aligned}\min_{\underline{L}} \sum_{(i,j)\in\mathcal{C}} \mathbb{E}\left[\left(\underline{L}^{\intercal}\underline{\widetilde{B}_{ij}} - B_{L,ij}\right)^2\right] \\ = \min_{\underline{L}} \underline{L}^{\intercal}\left(\sum_{(i,j)\in\mathcal{C}} \mathbb{E}\left[\underline{\widetilde{B}_{ij}}\,\underline{\widetilde{B}_{ij}}^{\intercal}\right]\right)\underline{L} - 2\left(\sum_{(i,j)\in\mathcal{C}} B_{L,ij}\underline{B_{ij}}^{\intercal}\right)\underline{L}\end{aligned} \tag{139}$$

which is associated to the optimality criterion

$$\left(\frac{1}{|\mathcal{C}|}\sum_{(i,j)\in\mathcal{C}} \mathbb{E}\left[\underline{\widetilde{B}_{ij}}\,\underline{\widetilde{B}_{ij}}^{\intercal}\right]\right)\underline{L} = \left(\frac{1}{|\mathcal{C}|}\sum_{(i,j)\in\mathcal{C}} B_{L,ij}\underline{B_{ij}}^{\intercal}\right) \tag{140}$$

Relating asymptotic quantities to expectations of sampled moments Using results from Ménétrier (2020, equation 4.2), we can relate asymptotic quantities to expectations of sample quantities:

$$\begin{aligned}B_{L,ij}B_{\ell,ij} = P_1(m)\,\mathbb{E}\left[\widehat{C}_m\left(X_{\ell,i}, X_{\ell,j}\right)\widehat{C}_m\left(X_{L,i}, X_{L,j}\right)\right] \\ + P_2(m)\left(\mathbb{E}\left[\widehat{C}_m\left(X_{\ell,i}, X_{L,i}\right)\widehat{C}_m\left(X_{\ell,j}, X_{L,j}\right)\right]\right. \\ \left. + \mathbb{E}\left[\widehat{C}_m\left(X_{\ell,i}, X_{L,j}\right)\widehat{C}_m\left(X_{\ell,j}, X_{L,i}\right)\right]\right) \\ + P_3(m)\,\mathbb{E}\left[\widehat{M}^4_m\left(X_{\ell,i}, X_{\ell,j}, X_{L,i}, X_{L,j}\right)\right]\end{aligned} \tag{141}$$

where $m \geq 4$ is a sampling size, $\widehat{C}_m$ is the unbiased Monte Carlo estimator of covariance for m samples, $\widehat{M}^4_m$ is the Monte Carlo estimator for fourth-order centered

moments, and $P_1(m)$, $P_2(m)$ and $P_3(m)$ are rational fractions.

$$\widehat{C}_m(X,Y) = \frac{1}{m-1}\sum_{i=1}^{m} \widetilde{X}^{(i)}\widetilde{Y}^{(i)} \tag{142}$$

$$\widehat{M}^4_m(X,Y,Z,T) = \frac{1}{m}\sum_{i=1}^{m} \widetilde{X}^{(i)}\widetilde{Y}^{(i)}\widetilde{Z}^{(i)}\widetilde{T}^{(i)} \tag{143}$$

$$\text{with} \quad \widetilde{X}^{(i)} = X^{(i)} - \widehat{E}_m(X) \quad \text{etc.} \tag{144}$$

$$P_1(m) = \frac{(m-1)(m^2-3m+1)}{m(m-2)(m-3)} \tag{145}$$

$$P_2(m) = \frac{m-1}{m(m-2)(m-3)} \tag{146}$$

$$P_3(m) = -\frac{m}{(m-2)(m-3)} \tag{147}$$

In practice, this relation requires estimating covariances between fine level L and any level ℓ. This can be done using a coupling group involving all fidelity levels. Ideally, the estimation should be performed independently of the covariance estimation.

Ergodicity assumption Using expression (141) to express the right-hand side of (140), we still have to express expectations of sample quantities. Estimating this expectations by single-sample Monte Carlo estimators is a possible solution. The high resulting sampling noise is actually cancelled out by the averaging over the whole equivalence class $\mathcal{C}$. Pretending that the noise is averaged out implicitly supposes that space averages are equivalent to sampling a common process (ergodic assumption). The approach of Benjamin Ménétrier presented in appendix D gives more insight on what this process may be.

Numerical cost In practice, the operator $\sum_{(i,j)\in\mathcal{C}} \mathbb{E}\left[\cdot\right]$ could also be replaced by an average over a random subset of $\mathcal{C}$ as is done in M15a.

Drawbacks In a single-level setting, Ménétrier and Auligné (2015) showed how the localization and hybridization weights could be jointly optimized. The approach was very appealing, but later trials showed it was less robust than a two-step optimization, where localization would be first optimized before optimizing the hybridization weights (Ménétrier, personal communication). A similar behavior may occur here, where having too many degrees of freedom in the optimization problem may result in non robust solutions.

A Retrieving the MLBLUE for the multidimensional expectation through constrained minimization of the variance

We retrieve here the results of section 4.5 with the variance minimization approach used in 3. The notations of section 4.5 are used here. In particular, all random quantities are column vectors.

We define the class of matrix-weighted ML estimators for $\boldsymbol{\alpha\mu} = \boldsymbol{\mu}_\alpha$ of the form

$$\widehat{\boldsymbol{\mu}}^{\mathrm{mat}} = \sum_{k=1}^{K} \sum_{\ell \in S^{(k)}} \boldsymbol{\beta}_\ell^{(k)} \widehat{E}^{(k)}\big[\mathbf{Z}_\ell\big] \tag{148}$$

where the $\boldsymbol{\beta}_\ell^{(k)}$ are matrices of $\mathbb{R}^{n \times n_\ell}$. The inner sum can be made implicit by joining these matrices into $\boldsymbol{\beta}^{(k)} := \big(\cdots\ \boldsymbol{\beta}_\ell^{(k)} \cdots\big)_{\ell \in S^{(k)}}$.

$$\widehat{\boldsymbol{\mu}}^{\mathrm{mat}} = \sum_{k=1}^{K} \boldsymbol{\beta}^{(k)} \widehat{E}^{(k)}\big[\mathbf{Z}^{(k)}\big] \tag{149}$$

Variance of the estimator The variance of the matrix-weighted estimator is given by

$$\operatorname{Tr} \operatorname{Cov}\big(\widehat{\boldsymbol{\mu}}^{\mathrm{mat}}, \widehat{\boldsymbol{\mu}}^{\mathrm{mat}}\big) = \sum_{k=1}^{K} \frac{1}{m^{(k)}} \boldsymbol{\beta}^{(k)} \mathbf{C}^{(k)} \left(\boldsymbol{\beta}^{(\boldsymbol{k})}\right)^{\intercal} \tag{150}$$

$$= \boldsymbol{\beta} \operatorname{Diag}_{k=1}^{K}\left(\frac{1}{m^{(k)}} \mathbf{C}^{(k)}\right) \boldsymbol{\beta}^{\intercal} \tag{151}$$

$$\text{with} \quad \mathbf{C}^{(k)} := \operatorname{Cov}\big(\mathbf{Z}^{(k)}, \mathbf{Z}^{(k)}\big) \tag{152}$$

$$\text{and} \quad \boldsymbol{\beta} := \big(\cdots\ \boldsymbol{\beta}^{(k)} \cdots\big)_{1 \le k \le K} \in \mathbf{R}^{n \times \sum_k n^{(k)}} \tag{153}$$

Unbiasedness constraint The no-bias condition is given by

$$\sum_{k=1}^{K} \boldsymbol{\beta}^{(k)} \mathbf{R}^{(k)} - \boldsymbol{\alpha} = 0 \tag{154}$$

$$\Leftrightarrow \qquad \boldsymbol{\beta} \mathbf{P}^{\intercal} - \boldsymbol{\alpha} = 0 \tag{155}$$

$$\text{with} \quad \mathbf{P} := \big(\mathbf{P}^{(1)} \ldots \mathbf{P}^{(K)}\big) \in \mathbf{R}^{N \times \sum_k n^{(k)}} \tag{156}$$

Unconstrained minimization problem We introduce Lagrange multipliers Λ.

$$\boldsymbol{\beta}, \boldsymbol{\Lambda} = \underset{\boldsymbol{\beta} \in \mathbf{R}^{p \times \sum_k n^{(k)}}, \boldsymbol{\Lambda} \in \mathbb{R}^{n \times N}}{\operatorname{argmin}} \frac{1}{2} \operatorname{Tr}\left(\boldsymbol{\beta} \operatorname{Diag}_{k=1}^{K}\left(\frac{1}{m^{(k)}} \mathbf{C}^{(k)} \right) \boldsymbol{\beta}^\intercal \right) - \sum_{i=1}^{n} \sum_{j=1}^{N} \Lambda_{ij} \left(\boldsymbol{\beta} \mathbf{P}^\intercal - \boldsymbol{\alpha} \right)_{ij} \quad (157)$$

Associated linear system Setting the gradient with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$ to zero yields the following linear system.

$$\boldsymbol{\beta} \boldsymbol{\Sigma} - \boldsymbol{\Lambda} \mathbf{P} = \mathbf{0} \quad (158)$$

$$\boldsymbol{\beta} \mathbf{P}^\intercal = \boldsymbol{\alpha} \quad (159)$$

with $\boldsymbol{\Sigma} := \operatorname{Diag}_{k=1}^{K}\left(\frac{1}{m^{(k)}} \mathbf{C}^{(k)} \right)$.

Optimal matrix weights The unique solution is given by

$$\boldsymbol{\beta} = \boldsymbol{\alpha} \boldsymbol{\phi}^{-1} \mathbf{P} \operatorname{Diag}_{k=1}^{K}\left(m^{(k)} \left(\mathbf{C}^{(k)} \right)^{-1} \right) \quad (160)$$

$$\text{with} \quad \boldsymbol{\phi} := \mathbf{P} \operatorname{Diag}_{k=1}^{K}\left(m^{(k)} \left(\mathbf{C}^{(k)} \right)^{-1} \right) \mathbf{P}^\intercal \quad (161)$$

$$= \sum_{k=1}^{K} m^{(k)} \mathbf{P}^{(k)} \left(\mathbf{C}^{(k)} \right)^{-1} \mathbf{R}^{(k)} \quad (162)$$

From which we can retrieve expressions (108) and (109) for $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\beta}_\ell^{(k)}$.

B Estimating the average covariance matrix of covariance estimators

The elements of the average covariance matrix $\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{C}^{(k,ij)}$ of covariance estimators, given by equation (125), are needed in two (not unrelated) situations:

1. To numerically optimize the weights and generalized sample allocation of the MLBLUE of a covariance matrix with scalar weights (see 6.2);

2. To numerically optimize the generalized sample allocation of an MLMC estimator of a covariance matrix (see appendix C).

The goal of this appendix is to provide computationally tractable ways to estimate the $L \times L$ matrix of these elements.

Each element of the matrix of interest can be decomposed into four terms. For $1 \leq k \leq K$ and $\ell, \ell' \in S^{(k)}$:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{C}_{\ell,\ell'}^{(k,ij)} = &\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\mathbb{M}^4\left[X_{\ell,i}, X_{\ell',i}, X_{\ell,j}, X_{\ell',j}\right]}{m^{(k)}} \\ &- \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\operatorname{Cov}\left(X_{\ell,i}, X_{\ell,j}\right)\operatorname{Cov}\left(X_{\ell',i}, X_{\ell',j}\right)}{m^{(k)}} \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\operatorname{Cov}\left(X_{\ell,i}, X_{\ell',j}\right)\operatorname{Cov}\left(X_{\ell,j}, X_{\ell',i}\right)}{m^{(k)}(m^{(k)}-1)} \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\operatorname{Cov}\left(X_{\ell,i}, X_{\ell',i}\right)\operatorname{Cov}\left(X_{\ell,j}, X_{\ell',j}\right)}{m^{(k)}(m^{(k)}-1)} \end{aligned} \tag{163}$$

As is, the double sums over the state dimension n prevent any explicit computation. Fortunately, naive Monte Carlo estimates of the centred moments can be reordered to make sure the cost of estimation stays linear in n.

Each one of the four terms can be estimated from (biased) Monte Carlo estimates based on coupled independent samples indexed by s, $1 \leq s \leq n_e$. We denote as $\widetilde{X}_{\ell',i}^{s} := X_{\ell,i}^{s} - \frac{1}{n_e}\sum_{s=1}^{n_e} X_{\ell,i}^{s}$ the centred perturbations associated to a sample $X_{\ell,i}^{s}$.

First term: fourth-order moment

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{M}^4\left[X_{\ell,i},X_{\ell',i},X_{\ell,j},X_{\ell',j}\right]\approx\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{n_e}\sum_{s=1}^{n_e}\widetilde{X}^s_{\ell,i}\widetilde{X}^s_{\ell',i}\widetilde{X}^s_{\ell,j}\widetilde{X}^s_{\ell',j} \tag{164}$$

$$=\frac{1}{n_e}\sum_{s=1}^{n_e}\left(\sum_{i=1}^{n}\widetilde{X}^s_{\ell,i}\widetilde{X}^s_{\ell',i}\right)^2 \tag{165}$$

Second term: product of intra-level covariances

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}\left(X_{\ell,i},X_{\ell,j}\right)\mathrm{Cov}\left(X_{\ell',i},X_{\ell',j}\right)\approx$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{n_e-1}\sum_{s=1}^{n_e}\widetilde{X}^s_{\ell,i}\widetilde{X}^s_{\ell,j}\right)\left(\frac{1}{n_e-1}\sum_{s'=1}^{n_e}\widetilde{X}^{s'}_{\ell',i}\widetilde{X}^{s'}_{\ell',j}\right)$$

$$=\frac{1}{(n_e-1)^2}\sum_{s=1}^{n_e}\sum_{s'=1}^{n_e}\left(\sum_{i=1}^{n}\widetilde{X}^s_{\ell,i}\widetilde{X}^{s'}_{\ell',i}\right)^2 \tag{166}$$

Third term: product of inter-level inter-point covariances

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}\left(X_{\ell,i},X_{\ell',j}\right)\mathrm{Cov}\left(X_{\ell,j},X_{\ell',i}\right)\approx$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{n_e-1}\sum_{s=1}^{n_e}\widetilde{X}^s_{\ell,i}\widetilde{X}^s_{\ell',j}\right)\left(\frac{1}{n_e-1}\sum_{s'=1}^{n_e}\widetilde{X}^{s'}_{\ell,j}\widetilde{X}^{s'}_{\ell',i}\right)$$

$$=\frac{1}{(n_e-1)^2}\sum_{s=1}^{n_e}\sum_{s'=1}^{n_e}\left(\sum_{i=1}^{n}\widetilde{X}^s_{\ell,i}\widetilde{X}^{s'}_{\ell',i}\right)\left(\sum_{i=1}^{n}\widetilde{X}^s_{\ell',i}\widetilde{X}^{s'}_{\ell,i}\right)$$

$$=\frac{1}{(n_e-1)^2}\sum_{s=1}^{n_e}\left(\sum_{i=1}^{n}\widetilde{X}^s_{\ell,i}\widetilde{X}^s_{\ell',i}\right)^2$$

$$+\frac{2}{(n_e-1)^2}\sum_{s=1}^{n_e}\sum_{s'>s}^{n_e}\left(\sum_{i=1}^{n}\widetilde{X}^s_{\ell,i}\widetilde{X}^{s'}_{\ell',i}\right)\left(\sum_{i=1}^{n}\widetilde{X}^s_{\ell',i}\widetilde{X}^{s'}_{\ell,i}\right) \tag{167}$$

Fourth term: product of inter-level covariances

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}\left(X_{\ell,i},X_{\ell',i}\right)\mathrm{Cov}\left(X_{\ell,j},X_{\ell',j}\right)\approx \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{n_e-1}\sum_{s=1}^{n_e}\widetilde{X}^s_{\ell,i}\widetilde{X}^s_{\ell',i}\right)\left(\frac{1}{n_e-1}\sum_{s'=1}^{n_e}\widetilde{X}^{s'}_{\ell,j}\widetilde{X}^{s'}_{\ell',j}\right) = \left(\frac{1}{n_e-1}\sum_{s=1}^{n_e}\sum_{i=1}^{n}\widetilde{X}^s_{\ell,i}\widetilde{X}^s_{\ell',i}\right)^2 \quad (168)$$

Computing in practice In practice, these quantities can be expressed as functions of the $n_e^2L^2$ space averages $\gamma(\ell,s,\ell',s')=\sum_{i=1}^{n}\widetilde{X}^s_{\ell,i}\widetilde{X}^{s'}_{\ell',i}$. Grouping the four terms together, we retrieve

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{C}^{(k,ij)}_{\ell,\ell'}=\frac{1}{m^{(k)}}\left(\frac{\sum_{s=1}^{n_e}\gamma(\ell,s,\ell',s)^2}{n_e}-\frac{\sum_{s=1}^{n_e}\sum_{s'=1}^{n_e}\gamma(\ell,s,\ell',s')^2}{(n_e-1)^2}\right) + \frac{1}{m^{(k)}(m^{(k)}-1)}\left(\frac{\sum_{s=1}^{n_e}\sum_{s'=1}^{n_e}\gamma(\ell,s,\ell',s')\gamma(\ell',s,\ell,s')}{(n_e-1)^2}+\frac{\left(\sum_{s=1}^{n_e}\gamma(\ell,s,\ell',s)\right)^2}{(n_e-1)^2}\right) \quad (169)$$

Note the symmetry $\gamma(\ell,s,\ell',s')=\gamma(\ell',s',\ell,s)$, which reduces to $n_e(n_e+1)L^2/2$ the number of space averages to compute.

A possible algorithm would be:

1. Generate n_e simulations coupled across all L levels: $\left((X^s_\ell)_{\ell=1}^{L}\right)_{s=1}^{n_e}$.
2. Loop over all ensemble members and all fidelity levels to estimate the L ensemble means $\mu_\ell\in\mathbb{R}^n$. These will be used to estimate fourth-order moments in next step.
3. Double loop over ensemble members and double loop over fidelity levels to estimate the point-wise γ. Make use of the symmetry property. Space-average.
4. Compute the elements of the averaged covariance matrix (169).

Remark This is valid in the limit of very large n_e. For finite sizes, these estimates are biased, as mentioned earlier. Unbiased estimators for these quantities do exist, but are more complex (see for instance Gerlovina and Hubbard, 2019).

C Optimal sample allocation for an MLMC covariance matrix estimator

The MLMC estimator for a covariance matrix is a specific (sub-optimal) case of (123), with weights $\beta^{(1)} = \begin{pmatrix}1\end{pmatrix}$ and $\beta^{(k)} = \begin{pmatrix}1 & -1\end{pmatrix}^\intercal$ for $2 \leq k \leq K$. The generalized variance of this multilevel estimator is given by (124), which can be estimated in practice following appendix B. Minimizing the variance as a function of the sample sizes solves the problem of optimal sample allocation. Note that the dependence on the sample sizes $\left(m^{(k)}\right)_{k=1}^K$ is hidden in the $\mathbb{C}^{(k,ij)}$. Note also that the minimization should be done numerically, since the variance involves complex terms such as the inverse of $m^{(k)}\left(m^{(k)}-1\right)$. This is quite direct in python, for instance using `scipy.optimize.minimize` with constraints.

Link with Mycek and De Lozzo (2019) A simpler but possibly suboptimal sample allocation can be found using the approach proposed by Mycek and De Lozzo (2019). They do not minimize the exact variance of the estimator, but an upper bound of this variance (their equation 2.31). Their bound is still written a a sum over the coupling groups, but each term is proportional to $1/m^{(k)}$. As a consequence, an analytical solution exists for the optimal sample allocation, which makes its determination much faster and simpler[1].

We extend the bound to the multivariate case, simply by summing over all i, j elements of the covariance matrix to be estimated. With our notations, the contributions to the variance include the sample size (different from the notations of Mycek and De Lozzo, 2019).

$$\mathcal{V}_k = \left(\beta^{(k)}\right)^\intercal \left(\sum_{i=1}^n \sum_{j=1}^n \mathbb{C}^{(k,ij)}\right) \beta^{(k)} \tag{170}$$

$$\mathcal{V}_k \leq \sum_{i=1}^n \sum_{j=1}^n \frac{1}{2} \frac{1}{m^{(k)}-1} \left[\sqrt{\mathbb{M}^4\left[\Delta_{\ell,i}\right] \mathbb{M}^4\left[\Sigma_{\ell,j}\right]} + \sqrt{\mathbb{M}^4\left[\Delta_{\ell,j}\right] \mathbb{M}^4\left[\Sigma_{\ell,i}\right]}\right] \tag{171}$$

$$= \frac{1}{m^{(k)}-1} \left(\sum_{i=1}^n \sqrt{\mathbb{M}^4\left[\Delta_{\ell,i}\right]}\right) \left(\sum_{i=1}^n \sqrt{\mathbb{M}^4\left[\Sigma_{\ell,i}\right]}\right) \tag{172}$$

where $\Delta_{\ell,i} = X_{\ell,i} - X_{\ell-1,i}$ and $\Sigma_{\ell,i} = X_{\ell,i} + X_{\ell-1,i}$ (assuming undefined variables are zero).

A possible algorithm would be:

[1] In the situations we tested, the bounds proposed by Mycek and De Lozzo (2019) were quite loose (about a factor 2 above the actual variances). This had no impact in the sample allocation though, as their relative evolution among levels was very similar to the true variances, which made them a useful proxy for the actual variance contributions.

1. Generate n_e simulations coupled across all L levels.

2. Loop over ensemble members and fidelity levels to estimate the L ensemble means $\mu_\ell \in \mathbb{R}^n$. These will be used to estimate fourth-order moments in next step.

3. Loop over ensemble members and over fidelity levels to estimate the pointwise fourth-order moments $\mathbb{M}^4\left[\Delta_{\ell,i}\right]$ and $\mathbb{M}^4\left[\Sigma_{\ell,i}\right]$.

4. Space-average and multiply to get the variance contribution terms.

D Optimal localization using random asymptotic quantities

This sections details how the results of section 6.4 can be derived (more rigorously?) with the formalism of M15a and M15b.

The main difference consists in considering asymptotic quantities as random, consistently with linear filtering theory. We assume the existence of two independent random processes $\mathcal{R}_1$ and $\mathcal{R}_2$. The first process generates asymptotic quantities denoted as $\mathcal{S}$ (for instance first, second and fourth order moments). The second one generates members consistent with these quantities (typically using an Ensemble of Data Assimilations).

The mean squared error is to be minimized over both process, though we only have access to one realization of $\mathcal{R}_1$. The expectation in the MSE is thus the expectation over both processes. We could either index the expectation operators by the processing they refer to, or use the formalism of conditional expectations:

$$\mathrm{MSE}\left(\widehat{\mathbf{B}}^{\mathrm{ML}}, \mathbf{B}_L\right) = \mathbb{E}\left[\left\|\sum_{k=1}^{K}\sum_{\ell\in S^{(k)}} \mathbf{L}_\ell^{(k)} \circ \widetilde{\mathbf{B}}_\ell^{(k)} - \mathbf{B}_L\right\|_{\mathrm{F}}^2\right] \tag{173}$$

$$= \mathbb{E}_1\left[\mathbb{E}_2\left[\left\|\sum_{k=1}^{K}\sum_{\ell\in S^{(k)}} \mathbf{L}_\ell^{(k)} \circ \widetilde{\mathbf{B}}_\ell^{(k)} - \mathbf{B}_L\right\|_{\mathrm{F}}^2\right]\right] \tag{174}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\left\|\sum_{k=1}^{K}\sum_{\ell\in S^{(k)}} \mathbf{L}_\ell^{(k)} \circ \widetilde{\mathbf{B}}_\ell^{(k)} - \mathbf{B}_L\right\|_{\mathrm{F}}^2\right]\middle|\,\mathcal{S}\right] \tag{175}$$

The rest of the derivation follows section 6.4.1. There is just one additional argument needed to simplify the cross terms $\mathbb{E}\left[B_L\underline{\widetilde{B}}^{\mathsf{T}}\right]$ when expanding the MSE:

Expectations of products of sampled and asymptotic quantities We have assumed the independence of the sampling error and the process $\mathcal{R}_2$ generating asymptotic statistics.

$$\mathbb{E}\left[\left(\widetilde{B}_\ell - B_\ell\right) B_L\right] = \mathbb{E}\left[\widetilde{B}_\ell - B_\ell\right]\mathbb{E}\left[B_L\right] \tag{176}$$

$$= 0 \tag{177}$$

$$i.e. \quad \mathbb{E}\left[\widetilde{B}_\ell B_L\right] = \mathbb{E}\left[B_\ell B_L\right] \tag{178}$$

So $\mathbb{E}\left[B_L\underline{\widetilde{B}}^{\mathsf{T}}\right] = \mathbb{E}\left[B_L\underline{B}\right]$.

Interpretation of the ergodicity hypothesis The ergodicity assumption intervenes within the same context of equivalence classes. A sum over a random subset of an equivalence class is meant to approximate the expectations over both random processes $\mathcal{R}_1$ and $\mathcal{R}_2$.

References

Bannister, R. N. (2017). "A review of operational methods of variational and ensemble-variational data assimilation." In: *Quarterly Journal of the Royal Meteorological Society* 143.703, pp. 607–633. DOI: 10.1002/qj.2982 (cit. on p. 3).

Buehner, M. (2005). "Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting." In: *Quarterly Journal of the Royal Meteorological Society* 131.607, pp. 1013–1043. DOI: 10.1256/qj.04.15 (cit. on p. 3).

Croci, M., K. E. Willcox, and S. J. Wright (2023). "Multi-output multilevel best linear unbiased estimators via semidefinite programming." In: arXiv: 2301.07831. URL: http://arxiv.org/abs/2301.07831 (cit. on pp. 2, 6, 11, 16, 17, 20, 23, 27).

Gerlovina, I. and A. E. Hubbard (2019). "Computer algebra and algorithms for unbiased moment estimation of arbitrary order." In: *Cogent Mathematics & Statistics* 6.1. Ed. by Y. Sun, p. 1701917. DOI: 10.1080/25742558.2019.1701917. eprint: https://doi.org/10.1080/25742558.2019.1701917. URL: https://doi.org/10.1080/25742558.2019.1701917 (cit. on p. 39).

Giles, M. B. (2008). "Multilevel Monte Carlo Path Simulation." In: *Operations Research* 56.3, pp. 607–617. DOI: 10.1287/opre.1070.0496 (cit. on p. 2).

— (2015). "Multilevel Monte Carlo methods." In: *Acta Numerica* 24, pp. 259–328. DOI: 10.1017/s096249291500001x (cit. on p. 2).

Gorodetsky, A. A., G. Geraci, M. S. Eldred, and J. D. Jakeman (2020). "A generalized approximate control variate framework for multifidelity uncertainty quantification." In: *Journal of Computational Physics* 408, p. 109257. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2020.109257. URL: https://www.sciencedirect.com/science/article/pii/S0021999120300310 (cit. on pp. 2, 7).

Lorenc, A. C. (2003). "The potential of the ensemble Kalman filter for NWP — a comparison with 4D-Var." In: *Quarterly Journal of the Royal Meteorological Society* 129.595, pp. 3183–3203. DOI: 10.1256/qj.02.132 (cit. on pp. 3, 30).

Ménétrier, B. (2020). *Sample covariance filtering*. DOI: 10.5281/ZENODO.4009099 (cit. on pp. 30, 33).

Ménétrier, B. and T. Auligné (2015). "Optimized localization and hybridization to filter ensemble-based covariances." In: *Monthly Weather Review* 143.10, pp. 3931–3947 (cit. on pp. 1, 34).

Ménétrier, B., T. Montmerle, Y. Michel, and L. Berre (2015a). "Linear Filtering of Sample Covariances for Ensemble-Based Data Assimilation. Part I: Optimality Criteria and Application to Variance Filtering and Covariance Localization." In: *Monthly Weather Review* 143.5, pp. 1622–1643. DOI: 10.1175/mwr-d-14-00157.1. URL: https://doi.org/10.1175/mwr-d-14-00157.1 (cit. on pp. 3, 26, 28, 30, 34, 42).

— (2015b). "Linear filtering of sample covariances for ensemble-based data assimilation. Part II: Application to a convective-scale NWP model." In: *Monthly Weather Review* 143.5, pp. 1644–1664. DOI: 10.1175/mwr-d-14-00156.1 (cit. on pp. 3, 30, 42).

Mycek, P. and M. De Lozzo (2019). "Multilevel Monte Carlo Covariance Estimation for the Computation of Sobol' Indices." In: *SIAM/ASA Journal on Uncertainty Quantification* 7.4, pp. 1323–1348. DOI: 10.1137/18m1216389 (cit. on pp. 27, 40).

Peherstorfer, B., K. Willcox, and M. Gunzburger (2018). "Survey of Multifidelity Methods in Uncertainty Propagation, Inference, and Optimization." In: *SIAM Rev.* 60.3, pp. 550–591. ISSN: 0036-1445. DOI: 10.1137/16M1082469. URL: https://doi.org/10.1137/16M1082469 (cit. on p. 2).

Schaden, D. and E. Ullmann (2020). "On Multilevel Best Linear Unbiased Estimators." In: *SIAM/ASA Journal on Uncertainty Quantification* 8.2, pp. 601–635. DOI: 10.1137/19m1263534 (cit. on pp. 1–7, 10, 20).

— (2021). "Asymptotic Analysis of Multilevel Best Linear Unbiased Estimators." In: *SIAM/ASA Journal on Uncertainty Quantification* 9.3, pp. 953–978. DOI: 10.1137/20m1321607 (cit. on pp. 2, 3, 6, 8, 10).