A generalized spectral decomposition technique to solve a class of linear stochastic partial differential equations

Anthony Nouy [a,*]

[a] *Institut de Recherche en Génie Civil et Mécanique (GeM), Nantes Atlantic University, Ecole Centrale Nantes, UMR CNRS 6183, 2 rue de la Houssinière, B.P. 92208, 44322 Nantes Cedex 3*

Abstract

We propose a new robust technique for solving a class of linear stochastic partial differential equations. The solution is approximated by a series of terms, each of which being the product of a scalar stochastic function by a deterministic function. None of these functions are fixed *a priori* but determined by solving a problem which can be interpreted as an "extended" eigenvalue problem. This technique generalizes the classical spectral decomposition, namely the Karhunen-Loève expansion. Ad-hoc iterative techniques to build the approximation, inspired by the power method for classical eigenproblems, then transform the problem into the resolution of a few uncoupled deterministic problems and stochastic equations. This method drastically reduces the calculation costs and memory requirements of classical resolution techniques used in the context of Galerkin stochastic finite element methods. Finally, this technique is particularly suitable to non-linear and evolution problems since it enables the construction of a relevant reduced basis of deterministic functions which can be efficiently reused for subsequent resolutions.

Key words: Computational Stochastic Mechanics, Stochastic Finite Element, Spectral Decomposition, Karhunen-Loève, Stochastic Partial Differential Equations

* Corresponding author. Tel.: +33(0)2-51-12-55-20; Fax: +33(0)2-51-12-52-52
Email address: anthony.nouy@univ-nantes.fr (Anthony Nouy).

Preprint

1 Introduction

Stochastic finite element methods have been recently proposed to solve stochastic partial differential equations and offer a significant tool to deal with the randomness which is inherent to mechanical systems.
Non-intrusive techniques, such as Monte-Carlo simulation [1,2], response surface method, projection [3] or regression [4] methods, have the great advantage that they only require the use of a simple deterministic calculation code. Stochastic problems, whatever their complexity, can be solved without any further developments, as long as the associated deterministic code exists. However, they require a huge number of deterministic calculations, which leads to high computational costs.
Galerkin-type methods [5–7], which differ from one another in the choice of the approximation space, systematically lead to a high precision solution which is explicit in terms of the basic random variables describing the uncertainties. However, they require the resolution of a huge system of equations. Ad-hoc Krylov-type iterative techniques have been proposed to make use of the sparsity of the system [8–10]. The difficulty to build efficient preconditioners and memory requirements induced by these techniques still limit their use to low stochastic dimensions when dealing with large scale applications.
In this paper, we propose a new alternative resolution technique to solve stochastic problems, inspired by a resolution technique for solving evolution equations [11,12]. The idea is to approximate the solution u by:

$$u \approx \sum_{i=1}^{M} \lambda_i U_i,$$

where the U_i are deterministic functions and where the λ_i are scalar stochastic functions (i.e. random variables). A decomposition of this type will be said optimal if the number of terms M is minimum for a given quality of approximation. The set of deterministic (resp. stochastic) functions can be then considered as an optimal deterministic (resp. stochastic) reduced basis. Here, neither the λ_i nor the U_i are fixed *a priori*. The key questions are then: how to define the "optimal" reduced basis and how to compute it? In fact, the obtained decomposition depends on what we mean by "optimal". If we knew the solution, the best approximation would classically be defined by minimizing the distance to the solution in a mean square sense. In this case, it simply leads to the classical spectral decomposition, namely the Karhunen-Loève expansion truncated at order M (see e.g. [5]). The problem is that the solution, and *a fortiori* its correlation, are not known. In [7], the authors propose to estimate the correlation of the solution by a Neumann expansion technique. Having the correlation, the deterministic functions are simply obtained by solving a classical eigenproblem. The associated random variables are finally computed by solving the initial problem on the reduced basis of deterministic

functions.
Here, we propose an intuitive and simple way to define the best decomposition, which generalizes the classical spectral decomposition. It leads to the resolution of a problem which can be interpreted as an "extended" eigenproblem. Ad-hoc iterative resolution techniques, inspired by classical techniques for solving eigenproblem, are then proposed. They transform the initial problem into the resolution of a few uncoupled deterministic problems and stochastic equations. This method then drastically reduces computational costs and memory requirements of classical resolution techniques.

The outline of the paper is as follows. In section 2, we introduce the abstract variational formulation of a class of linear stochastic partial differential equation, the discretization at the deterministic level and the stochastic modeling. In section 3, we describe the stochastic discretization and the classical Galerkin stochastic finite elements methods. Then, in section 4, we introduce the concept of generalized spectral decomposition. We will first focus on the case of symmetric problems before briefly introducing the case of non-symmetric problems. In section 5, we introduce power-type algorithms allowing to build the generalized spectral decomposition. Finally, in section 6, three examples will illustrate the efficiency of the proposed method.

2 Stochastic partial differential equation

2.1 Continuous problem

We first consider a deterministic partial differential equation which has the following variational formulation: find $u \in \mathcal{V}$ such that

$$a(u, v) = b(v) \quad \forall v \in \mathcal{V}, \tag{1}$$

where $\mathcal{V}$ is an appropriate space of admissible functions, a is a continuous bilinear form on $\mathcal{V}$ and b is a continous linear form on $\mathcal{V}$.
In the stochastic context, a and b forms are random. We denote by $(\Theta, \mathcal{B}, P)$ the probability space, where Θ is the set of outcomes, $\mathcal{B}$ the σ-algebra of events and P the probability measure. We denote the bilinear and linear forms respectively by $a(\cdot, \cdot; \theta)$ and $b(\cdot; \theta)$, which underlines their dependence on the outcome θ. The stochastic problem now consists in finding a stochastic process u which can be viewed as a random function with value in $\mathcal{V}$. Function space $\mathcal{V}$ can sometimes depend on the outcome, for example when dealing with random geometry [13]. Here, we consider that $\mathcal{V}$ does not depend on the outcome. The appropriate function space for u can then be chosen as the tensor product space $\mathcal{V} \otimes \mathcal{S}$, where $\mathcal{S}$ is an ad-hoc function space for real-valued random

functions. The weak formulation of the stochastic partial differential equation (see *e.g.* [14,15,7]) can then be written as follows: find $u \in \mathcal{V} \otimes \mathcal{S}$ such that

$$A(u, v) = B(v) \quad \forall v \in \mathcal{V} \otimes \mathcal{S}, \tag{2}$$

where

$$A(u, v) = \int_\Theta a(u(\theta), v(\theta); \theta) dP(\theta) := E(a(u, v)), \tag{3}$$

$$B(v) = \int_\Theta b(v(\theta); \theta) dP(\theta) := E(b(v)). \tag{4}$$

$E(\cdot)$ denotes the mathematical expectation. In this article, we focus on the case of a linear elliptic stochastic partial differential equation. The bilinear form A is continuous on $(\mathcal{V} \otimes \mathcal{S}) \times (\mathcal{V} \otimes \mathcal{S})$ and $(\mathcal{V} \otimes \mathcal{S})$-coercive, and the linear form B is continuous on $\mathcal{V} \otimes \mathcal{S}$. We consider that $\mathcal{S} = L^2(\Theta, dP)$ is an ad-hoc choice, such that $\mathcal{V} \otimes \mathcal{S} = \mathcal{V} \otimes L^2(\Theta, dP) \cong L^2(\Theta, dP; \mathcal{V})$ (see *e.g.* [14]).

2.2 Discretization at the deterministic level

The discretization at the deterministic level consists in searching an approximation of the solution of problem (1) under the form

$$u = \sum_{i=1}^{n} \varphi_i u_i(\theta), \tag{5}$$

where $u_i \in \mathcal{S}$ and where $\{\varphi_i\}_{i=1}^n$ is a basis of a finite dimensional space $\mathcal{V}_n \subset \mathcal{V}$. We will denote by $\mathbf{u} = (u_1, ..., u_n)^T \in \mathbb{R}^n \otimes \mathcal{S}$ the random vector of unknowns representing the approximate solution. It is the solution of the following semi-discretized problem: find $\mathbf{u} \in \mathbb{R}^n \otimes \mathcal{S}$ such that

$$E\left(\mathbf{v}^T \mathbf{A} \mathbf{u}\right) = E\left(\mathbf{v}^T \mathbf{b}\right) \quad \forall \mathbf{v} \in \mathbb{R}^n \otimes \mathcal{S}, \tag{6}$$

where $\mathbf{A} : \Theta \to \mathbb{R}^{n \times n}$ and $\mathbf{b} : \Theta \to \mathbb{R}^n$ are such that $\forall u, v \in \mathcal{V}_n$, we have P-almost surely

$$a(u, v; \theta) = \mathbf{v}^T \mathbf{A}(\theta) \mathbf{u}, \tag{7}$$

$$b(v; \theta) = \mathbf{v}^T \mathbf{b}(\theta). \tag{8}$$

Random matrix $\mathbf{A}$ and random vector $\mathbf{b}$ inherit respectively from continuity and coercivity properties of bilinear form A and continuity property of linear form B. Then, a solution in $\mathbb{R}^n \otimes \mathcal{S}$ of problem (6) exists and is unique.

We will then denote by $\|\cdot\|$ the classical norm on $\mathbb{R}^n \otimes \mathcal{S} \cong L^2(\Theta, dP; \mathbb{R}^n)$, defined by

$$\|\mathbf{u}\|^2 = E(\mathbf{u}^T \mathbf{u}). \tag{9}$$

We will denote by $((\mathbf{u}, \mathbf{v}))$ the associated scalar product.

2.3 Stochastic modeling

We consider that the probabilistic content of the problem is represented by a finite set of random variables $\boldsymbol{\xi} = (\xi_1, ..., \xi_m) : \Theta \longrightarrow \mathbb{R}^m$. This is naturally the case when A and B forms in (2) only depend on a finite set of parameters which are random variables. When these parameters are stochastic fields, an approximation step can allow their expression as functions of a finite set of random variables. For example, in the case of second-order random fields, such an approximation can consist of the truncated Karhunen-Loève expansion [5]. However, in the case of non-Gaussian random fields, the probabilistic characterization of the obtained finite set of random variables is not trivial. Some recent works try to answer this question by using truncated polynomial chaos expansions of the stochastic field, the coefficients of the expansions being constructed by the resolution of an adapted inverse problem (see *e.g.* [16,17]). Those techniques lead to a description of the probabilistic content by a finite set of independent Gaussian random variables. Let us note that an approximation is done on the bilinear form A. A particular care must then be taken on the truncation in order to keep coercivity properties of the approximate bilinear form [18,15] and then to ensure existence and uniqueness properties for the approximate model.

After this stochastic modeling step, a random function f can then be rewritten as a function of the basic random variables $f(\boldsymbol{\xi})$ and the stochastic problem (2) can be equivalently formulated on the finite dimensional probability space $(\Theta^{(m)}, \mathcal{B}^{(m)}, P^{(m)})$, where $\Theta^{(m)} = Range(\boldsymbol{\xi})$ is a subset of $\mathbb{R}^m$, $\mathcal{B}^m$ is the associated Borel σ-algebra and $P^{(m)}$ is the image probability measure (cf. [19,14]). In the following, for the sake of simplicity, we will still denote by $(\Theta, \mathcal{B}, P)$ the new finite dimensional probability space $(\Theta^{(m)}, \mathcal{B}^{(m)}, P^{(m)})$.

3 Classical stochastic finite element methods

3.1 Stochastic discretization

Classical stochastic finite element methods introduce a finite dimensional subspace $\mathcal{S}_P \subset \mathcal{S}$ for the approximation at the stochastic level. Several choices have been proposed for building approximation basis; spectral approaches (polynomial chaos [20,5], generalized chaos [21]) classically use orthogonal polynomial basis and show exponential convergence rates [21] in the case of

quite smooth solutions. For the case of non-smooth solutions, other approximation techniques have been introduced (Wiener-Haar chaos [22], finite elements [6]).
The approximation space is defined as follows:

$$\mathcal{S}_P = \{v \in \mathcal{S} \ ; \ v(\theta) = \sum_{\alpha \in \mathcal{I}_P} H_\alpha(\theta) v_\alpha, v_\alpha \in \mathbb{R}\}, \tag{10}$$

where $\{H_\alpha\}_{\alpha \in \mathcal{I}}$ is a basis of $\mathcal{S}$ and where $\mathcal{I}_P$ is a subset of $\mathcal{I}$ with cardinal P. If random variables ξ_i are independent, $\mathcal{S}$ is a tensor product space $\mathcal{S}^1 \otimes \ldots \otimes \mathcal{S}^m$. Each dimension can be independently discretized; denoting by $\alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{N}^m$ a multi-index, we can write $H_\alpha(\theta) = h^1_{\alpha_1}(\xi_1) \ldots h^m_{\alpha_m}(\xi_m)$, where $h^i_{\alpha_i} \in \mathcal{S}^i$. In practice, we will use an orthonormal basis, i.e. $E(H_\alpha H_\beta) = \delta_{\alpha\beta} = \delta_{\alpha_1\beta_1} ... \delta_{\alpha_m\beta_m}$.
In the case of mutually dependent random variables ξ_i, it is possible to use generalized "non-polynomial" chaos expansions [23] or more classically, to change the basic random variables by using an adapted mapping of the ξ_i into independent Gaussian random variables.
Below, we present a classical way of defining and computing an approximation of the solution of problem (6).

3.2 Galerkin approximation at the stochastic level

Classical Galerkin approximation of problem (6) is obtained by replacing the function space $\mathcal{V}_n \otimes \mathcal{S}$ by the approximation space $\mathcal{V}_n \otimes \mathcal{S}_P$. The problem is then to find $\mathbf{u} \in \mathbb{R}^n \otimes \mathcal{S}_P$ such that,

$$E\left(\mathbf{v}^T \mathbf{A} \mathbf{u}\right) = E\left(\mathbf{v}^T \mathbf{b}\right) \quad \forall \mathbf{v} \in \mathbb{R}^n \otimes \mathcal{S}_P, \tag{11}$$

which leads to the following system of equations:

$$\sum_{\beta \in \mathcal{I}_P} E(\mathbf{A} H_\alpha H_\beta) \mathbf{u}_\beta = E(H_\alpha \mathbf{b}) \quad \forall \alpha \in \mathcal{I}_P. \tag{12}$$

Denoting the solution by a block vector $\mathbf{u}$, whose block α is defined by $(\mathbf{u})_\alpha = \mathbf{u}_\alpha$, system (12) can be written

$$\mathbf{A}\mathbf{u} = \mathbf{b}, \tag{13}$$

where $\mathbf{A}$ and $\mathbf{b}$ are respectively a block matrix and a block vector. Their blocks are defined by $(\mathbf{A})_{\alpha\beta} = E(\mathbf{A} H_\alpha H_\beta)$ and $(\mathbf{b})_\alpha = E(H_\alpha \mathbf{b})$.

3.3 Krylov-type iterative techniques

In practice, system (13) can not be solved by a direct resolution technique. Indeed, memory requirements and computational costs of assembling and solving this huge system of size $P \times n$ become prohibitive for large-scale engineering problems. To avoid assembling and to take part of the sparsity of this system, we classically use a Krylov-type iterative resolution technique [8,9] such as preconditioned conjugate gradient (PCG) for symmetric problems, conjugate gradient square (CGS), etc. The resolution then only necessitates matrix-vector products which can be eventually parallelized [10]. The preconditioner $\mathbf{M}$ is classically taken as a block-diagonal matrix, each block being the inverse of the mean value of matrix $\mathbf{A}$, i.e. $(\mathbf{M})_{\alpha\beta} = E(\mathbf{A})^{-1}\delta_{\alpha\beta}$. The reason for this choice is that the preconditioner is quasi-optimal when $\mathbf{A}$ has low variance terms. As the variance of matrix $\mathbf{A}$ increases, the preconditioner $\mathbf{M}$ becomes less and less optimal. Iterative techniques then require a large number of iterations, which can drastically increase computational costs. Memory capacities required by these techniques can also be significant. Indeed, when dealing with such a huge system, reorthogonalization of the generated Krylov subspace is necessary, which implies the storage of this subspace. For example, if one considers a problem with $n = 10^5$ and $P = 5000$, storing the solution as double-precision floating-point numbers requires around 4 Gigabytes. Storing a Krylov subspace of dimension only 10 then requires 40 Gigabytes!

4 Generalized spectral decomposition

4.1 Principle

Here, we try to find an approximation of the solution of problem (11) under the form

$$\mathbf{u}(\theta) \approx \sum_{i=1}^{M} \lambda_i(\theta)\mathbf{U}_i \ , \tag{14}$$

where $\lambda_i \in \mathcal{S}_P$ are scalar random variables and $\mathbf{U}_i \in \mathbb{R}^n$ are deterministic vectors. A decomposition of this type will be said optimal if the number of terms M is minimum for a given quality of approximation. The set of deterministic vectors (resp. stochastic functions) would then be considered as an optimal deterministic (resp. stochastic) reduced basis. Neither the λ_i nor the $\mathbf{U}_i$ are fixed *a priori*. The key questions are then: how to define an "optimal" decomposition and how to compute it? The answer is of course related to what we mean by "optimal".

When we know the stochastic vector $\mathbf{u}$, a natural way to define the "best" approximation of the form (14) is to minimize the distance between the approximation and the solution:

$$\|\mathbf{u} - \sum_{i=1}^{M} \lambda_i \mathbf{U}_i\|^2 = \min_{\substack{\mathbf{U}_1,\ldots,\mathbf{U}_M \in \mathbb{R}^n \\ \lambda_1,\ldots,\lambda_M \in \mathcal{S}_P}} \|\mathbf{u} - \sum_{i=1}^{M} \lambda_i \mathbf{U}_i\|^2, \tag{15}$$

where $\|\cdot\|$ denotes the classical L^2-norm defined in (9). It is well known that the obtained approximation is the classical Karhunen-Loève expansion truncated at order M (cf. [5]). Vectors $\mathbf{U}_i$ are the M rightmost eigenvectors of $E(\mathbf{u}\mathbf{u}^T)$, which is the correlation matrix of $\mathbf{u}$, and can be characterized by

$$\begin{aligned} p(\mathbf{U}_i) &= \min_{V \in \mathbb{V}_{n-i+1}} \max_{\mathbf{U} \in V} \; p(\mathbf{U}), \\ \textit{with} \quad p(\mathbf{U}) &= \frac{\mathbf{U}^T E(\mathbf{u}\mathbf{u}^T)\mathbf{U}}{\mathbf{U}^T\mathbf{U}}, \end{aligned} \tag{16}$$

where p is the classical Rayleigh quotient and $\mathbb{V}_k$ is the set of all k-dimensional subspaces of $\mathbb{R}^n$. The associated λ_i are then obtained by $\lambda_i = (\mathbf{U}_i^T\mathbf{U}_i)^{-1}\mathbf{U}_i^T\mathbf{u}$.

In this section, we introduce an extension of this principle in order to build a decomposition of the solution of problem (11) without knowing this solution *a priori*. After an introduction of several notations and comments, we fill first focus on the case where operator $\mathbf{A}$ is symmetric before treating the general case.

Remark 1 *In this article, we consider that the approximation space $\mathcal{V}_n \otimes \mathcal{S}_P$ is given. The approximate solution $\mathbf{u}$ is then considered as our reference solution. For details on convergence properties of the approximation and estimation of errors with respect to the exact solution of* (2)*, see e.g. [14,15,24,25].*

4.2 Preliminaries and notations

In the following, we will denote by $\mathbf{W} = (\mathbf{U}_1 \ldots \mathbf{U}_M) \in \mathbb{R}^{n \times M}$ the matrix whose columns are the deterministic vectors and by $\mathbf{\Lambda} = (\lambda_1 \ldots \lambda_M)^T \in \mathbb{R}^M \otimes \mathcal{S}_P$ the stochastic vector whose components are the stochastic functions. The approximation of order M (14) will then be rewritten

$$\mathbf{u}^{(M)} = \sum_{i=1}^{M} \mathbf{U}_i \lambda_i = \mathbf{W}\mathbf{\Lambda}. \tag{17}$$

When neither $\mathbf{\Lambda}$ nor $\mathbf{W}$ are fixed, approximation (17) is not defined uniquely. In another words, there are infinitely many choices of stochastic functions and deterministic vectors leading to the same approximation. Indeed, for any

invertible matrix $\mathbf{P} \in \mathbb{R}^{M\times M}$, we clearly have

$$(\mathbf{WP})(\mathbf{P}^{-1}\mathbf{\Lambda}) = \mathbf{W}\mathbf{\Lambda}. \tag{18}$$

The couple composed by matrix ($\mathbf{WP}$) and stochastic vector ($\mathbf{P}^{-1}\mathbf{\Lambda}$) then yield the same approximation as the couple composed by matrix ($\mathbf{W}$) and stochastic vector ($\mathbf{\Lambda}$). Therefore, without loss of generality, we can for example impose orthogonality or orthonormality conditions on the λ_i or the $\mathbf{U}_i$ for the definition of the best approximation.

Finally, if the $\mathbf{U}_i$ (resp. the λ_i) were not linearly independent, it would be clearly possible to obtain a new decomposition of type (17) with a lower number of terms by using linear combinations of the $\mathbf{U}_i$ (resp. the λ_i).

Remark 2 *Random functions $\{\lambda_i\}_{i=1}^M$ are said "linearly independent" if they span a M-dimensional linear subspace of $\mathcal{S}_P$. In the finite dimensional space $\mathcal{S}_P$, the λ_i can be identified with vectors $\boldsymbol{\lambda}_i \in \mathbb{R}^P$ whose components are the coefficients $\lambda_{i,\alpha}$ of the λ_i on the basis $\{H_\alpha\}_{\alpha\in\mathcal{I}_P}$ of $\mathcal{S}_P$. The property "random variables $\{\lambda_i\}_{i=1}^M$ are linearly independent" is then equivalent to "vectors $\{\boldsymbol{\lambda}_i\}_{i=1}^M$ are linearly independent", or "the rank of matrix $(\boldsymbol{\lambda}_1 \dots \boldsymbol{\lambda}_M) \in \mathbb{R}^{P\times M}$ is M". Of course, it does not mean that random functions are statistically independent.*

We will suppose that (17) is the optimal decomposition which means that the $\mathbf{U}_i$ (resp. the λ_i) are linearly independent. We will denote by $\mathbb{G}_{n,M} = \{\mathbf{W} = (\mathbf{U}_1 \dots \mathbf{U}_M) \in \mathbb{R}^{n\times M}; rank(\mathbf{W}) = M\}$ the set of full rank matrices and by $\tilde{\mathbb{G}}_{P,M} = \{\mathbf{\Lambda} = (\lambda_1 \dots \lambda_M)^T \in \mathbb{R}^M \otimes \mathcal{S}_P \ ; \ dim(span(\{\lambda_i\}_{i=1}^M)) = M\}$ the set of linearly independent stochastic functions [1].

4.3 Case of a symmetric operator $\mathbf{A}$

In this section, we consider the particular case where the continuous coercive bilinear form A is also symmetric. Therefore, random matrix $\mathbf{A}$, which inherits from properties of A, is also symmetric. The right-hand side $\mathbf{b}$ is a stochastic vector.

[1] $span(\{\lambda_i\}_{i=1}^M)$ denotes the linear subspace of $\mathcal{S}_P$ which is spanned by the set of functions $\{\lambda_i\}_{i=1}^M$

4.3.1 Definition of the best approximation

The discretized problem (11) is equivalent to the following minimization problem:

$$\mathcal{J}(\mathbf{u}) = \min_{\mathbf{v}\in\mathbb{R}^n\otimes\mathcal{S}_P} \mathcal{J}(\mathbf{v}), \tag{19}$$

$$where \quad \mathcal{J}(\mathbf{v}) = E(\frac{1}{2}\mathbf{v}^T\mathbf{A}\mathbf{v} - \mathbf{v}^T\mathbf{b}). \tag{20}$$

A natural definition for the approximation follows:

Definition 3 *The best approximation of order M is defined by*

$$\mathcal{J}(\sum_{i=1}^{M}\lambda_i\mathbf{U}_i) = \min_{\substack{\mathbf{U}_1,\ldots,\mathbf{U}_M\in\mathbb{R}^n \\ \lambda_1,\ldots,\lambda_M\in\mathcal{S}_P}} \mathcal{J}(\sum_{i=1}^{M}\lambda_i\mathbf{U}_i), \tag{21}$$

which can be equivalently written in a matrix form:

$$\mathcal{J}(\mathbf{W}\mathbf{\Lambda}) = \min_{\substack{\mathbf{W}\in\mathbb{R}^{n\times M} \\ \mathbf{\Lambda}\in\mathbb{R}^M\otimes\mathcal{S}_P}} \mathcal{J}(\mathbf{W}\mathbf{\Lambda}). \tag{22}$$

4.3.2 Properties of the approximation

On one hand, the stationarity conditions of $\mathcal{J}(\mathbf{W}\mathbf{\Lambda})$ with respect to $\mathbf{\Lambda}$ writes: $\forall\mathbf{\Lambda}^* \in \mathbb{R}^M \otimes \mathcal{S}_P$,

$$E(\mathbf{\Lambda}^{*T}(\mathbf{W}^T\mathbf{A}\mathbf{W})\mathbf{\Lambda}) = E(\mathbf{\Lambda}^{*T}\mathbf{W}^T\mathbf{b}). \tag{23}$$

This is clearly the way we could have naturally defined the best stochastic functions associated with known deterministic vectors.
On the other hand, the stationarity conditions with respect to $\mathbf{W}$ writes: $\forall\mathbf{W}^* \in \mathbb{R}^{n\times M}$,

$$E(\mathbf{\Lambda}^T(\mathbf{W}^{*T}\mathbf{A}\mathbf{W})\mathbf{\Lambda}) = E(\mathbf{\Lambda}^T\mathbf{W}^{*T}\mathbf{b}). \tag{24}$$

This is still the way we could have naturally defined the best deterministic vectors associated with known stochastic functions. For example, if we impose the λ_i to be the basis functions of $\mathcal{S}_P$, namely the H_α, equation (24) would simply yields the classical solution of system (11), i.e. $\mathbf{W} = (\ldots\mathbf{u}_\alpha\ldots)$. However, the resulting approximation is the less optimal approximation since it has the maximum number of terms $M = P$.

Here, we ask the best approximation for simultaneously verifying equations (23) and (24). The following proposition gives fundamental properties allowing to better understand the meaning of the approximation and to develop computational resolution techniques.

Proposition 4 *The best approximation defined by definition 3 is characterized by:*

- $\mathbf{W}$ *maximizes on the set of full rank matrices* $\mathbb{G}_{n,M}$ *the functional* $R(\mathbf{W})$ *defined by*

$$\begin{aligned} & R(\mathbf{W}) = Trace(\mathbf{R}(\mathbf{W})), \\ with \quad & \mathbf{R}(\mathbf{W}) = E((\mathbf{W}^T\mathbf{A}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{b}\mathbf{b}^T\mathbf{W})). \end{aligned} \tag{25}$$

- *The stochastic functions are obtained by*

$$\mathbf{\Lambda} = (\mathbf{W}^T\mathbf{A}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{b}. \tag{26}$$

Moreover, the best approximation verifies

$$\mathcal{J}(\mathbf{W}\mathbf{\Lambda}) = -\frac{1}{2}R(\mathbf{W}). \tag{27}$$

Proof. Equation (23) yields relation (26). Then, we have

$$\begin{aligned} \mathcal{J}(\mathbf{\Lambda}\mathbf{W}) &= \frac{1}{2}E(\mathbf{\Lambda}^T\mathbf{W}^T\mathbf{A}\mathbf{W}\mathbf{\Lambda}) - E(\mathbf{\Lambda}^T\mathbf{W}^T\mathbf{b}) \\ &= -\frac{1}{2}E(\mathbf{\Lambda}^T\mathbf{W}^T\mathbf{b}) \\ &= -\frac{1}{2}E(\mathbf{b}^T\mathbf{W}(\mathbf{W}^T\mathbf{A}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{b}) \\ &= -\frac{1}{2}Trace(E((\mathbf{W}^T\mathbf{A}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{b}\mathbf{b}^T\mathbf{W}))) \\ &= -\frac{1}{2}R(\mathbf{W}). \end{aligned}$$

The best deterministic vectors are then such that $\mathbf{W} \in \mathbb{G}_{n,M}$ maximizes $R(\mathbf{W})$. □

Remark 5 *In equation* (26)*, we use an abuse of notation. Indeed, the quantity* $(\mathbf{W}^T\mathbf{A}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{b}$ *does not necessarily belongs to* $\mathbb{R}^M \otimes \mathcal{S}_P$*. In fact,* $\mathbf{\Lambda} \in \mathbb{R}^M \otimes \mathcal{S}_P$ *must be interpreted as the solution of problem* (23)*. This abuse of notation is also made in the definition of functional* $R(\mathbf{W})$*. In fact, it should be interpreted as follows:* $R(\mathbf{W}) = Trace(E(\mathbf{\Lambda}^T\mathbf{b}^T\mathbf{W}))$ *where* $\mathbf{\Lambda}$ *is the solution of* (23)*.*

For all invertible matrix $\mathbf{P} \in \mathbb{R}^{M\times M}$,

$$\mathbf{R}(\mathbf{W}\mathbf{P}) = \mathbf{P}^{-1}\mathbf{R}(\mathbf{W})\mathbf{P}, \tag{28}$$

and

$$R(\mathbf{W}\mathbf{P}) = R(\mathbf{W}). \tag{29}$$

Equation (29) is a property of homogeneity. In another words, functional R takes the same value for all matrices whose column vectors span a given M-dimensional linear subspace of $\mathbb{R}^n$. This is related to the non-uniqueness of the best solution $\mathbf{W}$ (cf. section (4.2)). For the decomposition to be defined uniquely, we could impose orthonormality conditions on $\mathbf{W}$. For example, denoting by $\mathbf{M}$ a symmetric definite positive matrix, the optimization problem on R could be defined on the space of $\mathbf{M}$-orthogonal matrices $\mathbb{G}^*_{n,M} = \{\mathbf{W} \in \mathbb{G}_{n,M}; \mathbf{W}^T\mathbf{M}\mathbf{W} = \mathbf{I}_M\}$, where $\mathbf{I}_M$ is the identity matrix on $\mathbb{R}^M$. The set $\mathbb{G}^*_{n,M}$ is called a Stiefel manifold (cf. [26]).

4.3.3 Case of deterministic operator $\mathbf{A}$

In order to interpret the approximation defined in proposition 4, let us consider the particular case where operator $\mathbf{A}$ is deterministic. In this case, we have

$$\mathbf{R}(\mathbf{W}) = (\mathbf{W}^T\mathbf{A}\mathbf{W})^{-1}(\mathbf{W}^T E(\mathbf{b}\mathbf{b}^T)\mathbf{W}). \tag{30}$$

The stationarity condition of $R(\mathbf{W}) = Trace(\mathbf{R}(\mathbf{W}))$ writes

$$\mathbf{A}\mathbf{W}\mathbf{R}(\mathbf{W}) = E(\mathbf{b}\mathbf{b}^T)\mathbf{W}, \tag{31}$$

which is a classical generalized eigenproblem written in a matrix form. $\mathbf{R}(\mathbf{W})$ (resp. $R(\mathbf{W})$) is the associated matrix (resp. scalar) Rayleigh quotient. Therefore, the best $\mathbf{W}$ characterized in proposition 4 is such that its column vectors span the rightmost M-dimensional eigenspace of the generalized eigenproblem (31). A particular choice for the columns of $\mathbf{W} = (\mathbf{U}_1 \dots \mathbf{U}_M)$ consists of the M rightmost eigenvectors. The $\mathbf{U}_i$ are $\mathbf{A}$-orthogonal and the associated λ_i are characterized by

$$\lambda_i = (\mathbf{U}_i^T\mathbf{A}\mathbf{U}_i)^{-1}\mathbf{U}_i^T\mathbf{b}. \tag{32}$$

The couples $(\mathbf{U}_i, \lambda_i)$ also verify

$$\mathcal{J}(\sum_{i=1}^{M} \lambda_i \mathbf{U}_i) = -\frac{1}{2}\sum_{i=1}^{M} R(\mathbf{U}_i). \tag{33}$$

$E(\mathbf{b}\mathbf{b}^T)$ is the correlation matrix of random vector $\mathbf{b}$. The spectral decomposition can then simply be interpreted as a truncated Karhunen-Loève expansion of $\mathbf{b}$ in the metric induced by $\mathbf{A}$. Here, we emphasize that the approximation is not unique. Infinitely many choices of $\mathbf{W}$ yield the same approximation as the particular choice consisting of eigenvectors.

Remark 6 *If we choose for the* $\mathbf{U}_i$ *the eigenvectors of the generalized eigenproblem* (31)*, which are* $\mathbf{A}$*-orthogonal, the associated stochastic functions are orthogonal. We have then simultaneous orthogonality properties for both the stochastic functions and the deterministic vectors. In general, this property can not be verified in the case of a stochastic operator* $\mathbf{A}$*.*

4.3.4 Interpretation and comments

Regarding the results of the previous section, functional $\mathbf{R}(\mathbf{W})$ (resp. $R(\mathbf{W})$), defined in proposition 4, will be called a "generalized" matrix (resp. scalar) Rayleigh quotient. In particular, R has the homogeneity property (29) of a classical Rayleigh quotient. The best approximation obtained will also be called a generalized spectral decomposition, although there is no associated classical eigenproblem.
One could think that since the obtained decomposition is not the Karhunen-Loève expansion, it is not the optimal decomposition. In fact, the obtained decomposition is optimal with respect to the optimality criterium introduced in definition 3. Properties of random matrix $\mathbf{A}$, inherited from the continuous coercive bilinear form A, allow to define the following norm on $\mathbb{R}^n \otimes \mathcal{S}_P$:

$$\|\mathbf{v}\|_{\mathbf{A}}^2 = E\left(\mathbf{v}^T \mathbf{A} \mathbf{v}\right). \tag{34}$$

We can then easily show that the decomposition characterized by proposition 4 verifies

$$\|\mathbf{u} - \mathbf{W}\mathbf{\Lambda}\|_{\mathbf{A}}^2 = \|\mathbf{u}\|_{\mathbf{A}}^2 - \|\mathbf{W}\mathbf{\Lambda}\|_{\mathbf{A}}^2 = \|\mathbf{u}\|_{\mathbf{A}}^2 - R(\mathbf{W}). \tag{35}$$

The obtained decomposition is then optimal with respect to the $\mathbf{A}$-norm (34) while the direct Karhunen-Loève expansion of the solution is optimal with respect to the L^2-norm (9).

As a last comment, let us mention that it is also possible to characterize the best approximation by defining an optimization problem on the stochastic functions. We can show that the following proposition 7 is equivalent to proposition 4. For this new proposition, we use the following notations: we denote by $E(\mathbf{\Lambda} \otimes \mathbf{A} \otimes \mathbf{\Lambda}) \in \mathbb{R}^{M \times n \times n \times M}$ the four indices matrix such that $(E(\mathbf{\Lambda} \otimes \mathbf{A} \otimes \mathbf{\Lambda}))_{ijkl} = E(\lambda_i A_{jk} \lambda_l)$. We denote by $E(\mathbf{b} \otimes \mathbf{\Lambda}) \in \mathbb{R}^{n \times M}$ the matrix such that $(E(\mathbf{b} \otimes \mathbf{\Lambda}))_{ij} = E(b_i \lambda_j)$. We then define the operation ":" as follows: for $\mathbf{W}, \mathbf{W}^* \in \mathbb{R}^{n \times M}$,

$$\begin{aligned} \mathbf{W}^{*T} : E(\mathbf{\Lambda} \otimes \mathbf{A} \otimes \mathbf{\Lambda}) : \mathbf{W} &= \sum_{i,j,k,l} W_{ji}^* E(\lambda_i A_{jk} \lambda_l) W_{kl} \\ &= E(\mathbf{\Lambda}^T \mathbf{W}^{*T} \mathbf{A} \mathbf{W} \mathbf{\Lambda}). \end{aligned}$$

Proposition 7 *The best approximation defined by definition 3 is characterized by:*

- $\mathbf{\Lambda}$ *maximizes on* $\widetilde{\mathbb{G}}_{P,M}$ *the functional* $\widetilde{R}(\mathbf{\Lambda})$ *defined by*

$$\widetilde{R}(\mathbf{\Lambda}) = E(\mathbf{\Lambda} \otimes \mathbf{b}) : E(\mathbf{\Lambda} \otimes \mathbf{A} \otimes \mathbf{\Lambda})^{-1} : E(\mathbf{b} \otimes \mathbf{\Lambda}). \tag{36}$$

- *The deterministic vectors are obtained by*

$$\mathbf{W}^T = E(\mathbf{\Lambda} \otimes \mathbf{A} \otimes \mathbf{\Lambda})^{-1} : E(\mathbf{b} \otimes \mathbf{\Lambda}). \tag{37}$$

Moreover, the best approximation verifies

$$\mathcal{J}(\mathbf{W}\mathbf{\Lambda}) = -\frac{1}{2}\widetilde{R}(\mathbf{\Lambda}). \tag{38}$$

4.3.5 Another definition of the approximation

The best approximation can also be naturally defined by the following optimization problems, which define the couples $(\lambda_i, \mathbf{U}_i)$ one after the other:

Definition 8 *The best approximation of order M can be defined recursively: for $i = 1, ..., M$*

$$\mathcal{J}(\lambda_i \mathbf{U}_i) = \min_{\mathbf{U}\in\mathbb{R}^n, \lambda\in\mathcal{S}_P} \mathcal{J}(\lambda\mathbf{U} + \sum_{j=1}^{i-1} \lambda_j \mathbf{U}_j) \tag{39}$$

Following the proof of proposition 4, we obtain the following characterization of the new obtained approximation.

Proposition 9 *The approximation defined in definition 8 can be charaterized by* [2]

- $\mathbf{U}_i$ *maximizes the generalized Rayleigh quotient*

$$R_i(\mathbf{U}) = E(\mathbf{U}^T\mathbf{b}_i(\mathbf{U}^T\mathbf{A}\mathbf{U})^{-1}\mathbf{b}_i^T\mathbf{U})$$
$$with \quad \mathbf{b}_i = \mathbf{b} - \sum_{j=1}^{i-1} \mathbf{A}\lambda_j\mathbf{U}_j$$

- $\lambda_i = (\mathbf{U}_i^T\mathbf{A}\mathbf{U}_i)^{-1}\mathbf{U}_i^T\mathbf{b}_i$.

Moreover, the best approximation verifies

$$\mathcal{J}(\sum_{i=1}^{M} \lambda_i\mathbf{U}_i) = -\frac{1}{2}\sum_{i=1}^{M} R_i(\mathbf{U}_i). \tag{40}$$

In the case of a deterministic operator, we can easily show that definitions 8 and 3 yield the same approximation. However, these definitions do not match in general in the case of a stochastic operator. Definition 3 then clearly yields a better approximation as definition 8. The latter definition can however be interesting from a computational point of view, as we will see in section 5.

[2] We use the same abuse of notation as in proposition 4, explained in Remark 5

4.4 Case of a non-symmetric operator

For the case of non-symmetric operator, there is no direct minimization problem associated with the variational formulation (6). A possible and natural way to define an approximation $\mathbf{W}\mathbf{\Lambda}$ is to write the two orthogonality criteria (23) and (24) which can be obtained by introducing in (6) test functions of the form $\mathbf{v} = \mathbf{W}\mathbf{\Lambda}^* + \mathbf{W}^*\mathbf{\Lambda}$. In the case of a non-symmetric deterministic operator $\mathbf{A}$, we can easily show that it leads to a classical non-symmetric generalized eigenproblem. Therefore, the approximation still has a full meaning but no characterization as the one in proposition 4 can be derived.

Another idea consists in reformulating the problem as a minimization problem in a least-square sense. Let us denote by $\mathbf{M} \in \mathbb{R}^{n\times n}$ a symmetric positive definite matrix which defines the following scalar product on $\mathbb{R}^n \otimes \mathcal{S}_P$:

$$((\mathbf{U}, \mathbf{V}))_{\mathbf{M}} = E(\mathbf{U}^T\mathbf{M}\mathbf{V}). \tag{41}$$

We will denote by $\|\cdot\|_{\mathbf{M}}$ the associated norm. For example, if the expected value of the symmetric part of $\mathbf{A}$ is positive definite, we can take $\mathbf{M} = (\frac{1}{2}E(\mathbf{A}+\mathbf{A}^T))^{-1}$. We can then formulate the problem as a minimization problem of the norm of the residual:

$$\begin{aligned} &\mathcal{E}(\mathbf{u}) = \min_{\mathbf{v}\in\mathbb{R}^n\otimes\mathcal{S}_P} \mathcal{E}(\mathbf{v}) \\ &where \quad \mathcal{E}(\mathbf{v}) = \|\mathbf{b} - \mathbf{A}\mathbf{v}\|^2_{\mathbf{M}}. \end{aligned} \tag{42}$$

The previous theoretical results associated with symmetric operators can then be still applied by replacing operator $\mathbf{A}$ by $\mathbf{A}^T\mathbf{M}\mathbf{A}$ and right hand side $\mathbf{b}$ by $\mathbf{A}^T\mathbf{M}\mathbf{b}$. Of course, the approximate solution of this problem is not in general the same as for problem (11) when $\mathbf{A}$ is random.

Although this symmetrization is well adapted to the discretized formulation (finite dimensional framework), it is not easy to transpose into the continuous framework. Indeed, the introduction of adjoint operator is non trivial, essentially due to the treatment of boundary conditions. Moreover, ad-hoc function spaces need for more regularity after symmetrization and the building of approximation spaces is then more difficult.

However, we will see that this symmetrization does not seem necessary since algorithms which are given in the following section also gives satisfactory results when directly applied to non-symmetric problems. Without the symmetry, we do not have an optimality criterium as in proposition 4. It is then more difficult to judge the quality of the approximation. Of course, it is possible to evaluate the norm of the residual of equation (11) but this increases computational costs.

5 Power-type algorithms for the construction of the spectral decomposition

5.1 Description of the algorithm

Here, we propose a first algorithm to build the spectral decomposition defined in definition 8. The couples $(\lambda_i, \mathbf{U}_i)$ are built one after the other. Each couple must solve a coupled minimization problem of the functional $\mathcal{J}$. A natural and simple idea to build the approximation is to minimize $\mathcal{J}$ alternatively on the stochastic function and on the deterministic vector. Algorithm 1, named (P-GSD) for Power-type Generalized Spectral Decomposition, follows this idea. Some trivial computational improvements are straightforward (redundancy of several operations).

Algorithm 1 *Power-type Generalized Spectral Decomposition (P-GSD)*

1: $\mathbf{u} := 0$, $\tilde{\mathbf{b}} := \mathbf{b}$
2: ***for*** $i = 1$ *to* M ***do***
3: $\quad \lambda_i := \lambda_0$, $R_i^{(0)} := 0$
4: $\quad$ ***for*** $k = 1$ *to* k_{max} ***do***
5: $\quad\quad \mathbf{U}_i := E(\mathbf{A}\lambda_i^2)^{-1}E(\tilde{\mathbf{b}}\lambda_i)$
6: $\quad\quad \mathbf{U}_i := \mathbf{U}_i/\|\mathbf{U}_i\|_{\mathbf{M}}$
7: $\quad\quad \lambda_i = (\mathbf{U}_i^T\mathbf{A}\mathbf{U}_i)^{-1}\mathbf{U}_i^T\tilde{\mathbf{b}}$
8: $\quad\quad R_i^{(k)} := E(\mathbf{U}_i^T\tilde{\mathbf{b}}\lambda_i)$
9: $\quad\quad \gamma := \left|R_i^{(k)} - R_i^{(k-1)}\right| / R_i^{(k)}$
10: $\quad\quad$ ***if*** $\gamma < \gamma_{stop}$ ***then***
11: $\quad\quad\quad$ *break*
12: $\quad\quad$ ***end if***
13: $\quad$ ***end for***
14: $\quad \mathbf{u} := \mathbf{u} + \lambda_i\mathbf{U}_i$
15: $\quad \tilde{\mathbf{b}} := \tilde{\mathbf{b}} - \mathbf{A}\lambda_i\mathbf{U}_i$
16: $\quad$ *Compute error indicator* $\epsilon^{(i)}$
17: $\quad$ ***if*** $\epsilon^{(i)} < \epsilon_{stop}$ ***then***
18: $\quad\quad$ *break*
19: $\quad$ ***end if***
20: ***end for***

Let us explain this algorithm. Step 1 is the initialization step of the solution and the residual. The loop beginning at step 2 corresponds to the recursive building of couples $(\lambda_i, \mathbf{U}_i)$. At step 3, we initialize the stochastic function $\lambda_i = \lambda_0 \in \mathcal{S}_P$. We will simply always take $\lambda_{0,\alpha} = 1$, $\forall \alpha \in \mathcal{I}_P$. We will see that this very simple initialization gives satisfactory results in practice. The loop beginning at step 4 is the alternate minimization procedure. For a fixed stochastic function λ_i, the stationarity condition with respect to $\mathbf{U}_i$ writes

$E(\mathbf{A}\lambda_i^2)\mathbf{U}_i = E(\tilde{\mathbf{b}}\lambda_i)$. It has to be noticed that it is a simple deterministic problem whose resolution is relatively cheap. This system is solved at step 5 and $\mathbf{U}_i$ is normalized at step 6. In the symmetric case, we simply take $\mathbf{M} = E(\mathbf{A})$ for the definition of the norm. Then, for a fixed $\mathbf{U}_i$, the stationarity condition with respect to λ_i writes

$$E(\lambda^* \mathbf{U}_i^T \mathbf{A} \mathbf{U}_i \lambda_i) = E(\lambda^* \mathbf{U}_i^T \tilde{\mathbf{b}}) \quad \forall \lambda^* \in \mathcal{S}_P. \tag{43}$$

This system is solved at step 7. From step 8 to step 12, we introduce a stopping criterium based on the convergence of the quantity $R_i^{(k)}$. This quantity corresponds to the generalized Rayleigh quotient in the case where operator $\mathbf{A}$ is symmetric (see definition 8). At steps 14 and 15, we reactualize the solution and the residual. From 16 to 19, we introduce a stopping criterium. The error indicator can be the residual error

$$\epsilon_{res}^{(i)} = \frac{\|\mathbf{b} - \mathbf{A} \sum_{j=1}^{i} \lambda_j \mathbf{U}_j\|}{\|\mathbf{b}\|}. \tag{44}$$

Regarding the results of proposition 9, we can also use the following error indicator, based on the generalized Rayleigh quotient evaluation:

$$\epsilon_{ray}^{(i)} = \frac{R_i(\mathbf{U}_i)}{\sum_{j=1}^{i} R_j(\mathbf{U}_j)}. \tag{45}$$

This last indicator has the advantage to be cheaper to compute. Indicators (44) and (45) will be compared in section 6.

Remark 10 *Computing the matrix and right-hand side of the deterministic problem (step 5) requires the computation of quantities such as $E(\mathbf{A}\lambda_i\lambda_j)$ or $E(\mathbf{b}\lambda_i)$. This kind of computations are classical within the context of stochastic finite element methods. Let us consider that the random vector $\mathbf{b}$ is decomposed as follows: $\mathbf{b} = \sum_{k=1}^{M_b} b_k(\theta)\mathbf{b}_k$, with $b_k(\theta) = \sum_{\alpha \in \mathcal{I}_P} b_{k,\alpha} H_\alpha(\theta) \in \mathcal{S}_P$. Then, due to orthonormality property of the basis functions H_α, $E(\mathbf{b}\lambda_i) = \sum_{k=1}^{M_b} \mathbf{b}_k \sum_{\alpha \in \mathcal{I}_P} b_{k,\alpha} \lambda_{i,\alpha}$. Let us now consider that the random matrix writes as follows: $\mathbf{A} = \sum_{k=1}^{M_A} a_k(\theta)\mathbf{A}_k$, where the a_k are random variables. Then $E(\mathbf{A}\lambda_i\lambda_j) = \sum_{k=1}^{M_A} \mathbf{A}_k E(a_k \lambda_i \lambda_j)$. In practice, one pre-computes and stores the matrices $\mathbf{\Delta}^{(k)}$ whose components are $(\mathbf{\Delta}^{(k)})_{\alpha\beta} = E(a_k H_\alpha H_\beta)$ and such that $E(a_k\lambda_i\lambda_j) = \sum_{\alpha,\beta \in \mathcal{I}_P} (\mathbf{\Delta}^{(k)})_{\alpha\beta} \lambda_{i,\alpha}\lambda_{j,\beta}$. If the a_k are decomposed on the stochastic basis, i.e. $a_k(\theta) = \sum_\gamma a_{k,\gamma} H_\gamma(\theta)$, then $(\mathbf{\Delta}^{(k)})_{\alpha\beta} = \sum_\gamma a_{k,\gamma} E(H_\gamma H_\alpha H_\beta)$, where $E(H_\gamma H_\alpha H_\beta)$ only depends on the chosen basis functions $\{H_\alpha\}$.*

5.2 Interpretation and comments

In the case of deterministic operator, iteration k of the alternate minimization stage consists in reactualizing the deterministic vector in the following

way: $\mathbf{U}_i \leftarrow \mathbf{A}^{-1}E(\tilde{\mathbf{b}}\tilde{\mathbf{b}}^T)\mathbf{U}_i/\eta$, where η is a normalizing scalar. The proposed algorithm is then equivalent to a classical power method to solve the associated generalized eigenproblem $\mathbf{A}\mathbf{U}_i = \eta E(\tilde{\mathbf{b}}\tilde{\mathbf{b}}^T)\mathbf{U}_i$. It classically converges toward the rightmost eigenvector. We will see in examples that this algorithm is also efficient in the case of stochastic operator. In practice, the alternate minimization procedure (steps 4 to 13) converges very fast. We will then classically limit the number of iterations k_{max} to 3 or 4. In the case of deterministic non-symmetric operator, we know that the proposed power algorithm is not always convergent. If the maximum amplitude eigenvalue is complex, the generalized eigenproblem admits 2 complex conjugate eigenvalues. Vector $\mathbf{U}_i$ does not converge in this case. However, it tends to stay in the subspace generated by the real and complex parts of the associated complex eigenvectors and the obtained couple $(\lambda_i, \mathbf{U}_i)$ still happens to be pertinent. We will see in examples that this algorithm also gives satisfactory results in the general case of random eventually non-symmetric operators.

5.3 Power-type algorithm with updating

Let us suppose we have built an approximation $\sum_{i=1}^{M} \lambda_i \mathbf{U}_i = \mathbf{W}\mathbf{\Lambda}$. We have seen that optimality depends on the way we define the "best" decomposition. Indeed, definitions 3 and 8 match only in the case where operator $\mathbf{A}$ is deterministic. Once we have obtained such a decomposition, it can then be interesting to update the decomposition. A natural way to do this is to fix the deterministic vectors $\mathbf{U}_i$ and to compute new stochastic functions $\lambda_i \in \mathcal{S}_P$ by using a Galerkin orthogonality criterium (23). The updating can then be formulated: find $\mathbf{\Lambda} \in \mathbb{R}^M \otimes \mathcal{S}_P$ such that $\forall \mathbf{\Lambda}^* \in \mathbb{R}^M \otimes \mathcal{S}_P$,

$$E(\mathbf{\Lambda}^{*T}(\mathbf{W}^T\mathbf{A}\mathbf{W})\mathbf{\Lambda}) = E(\mathbf{\Lambda}^{*T}\mathbf{W}^T\mathbf{b}). \tag{46}$$

To obtain the solution $\mathbf{\Lambda}(\theta) = \sum_{\alpha\in\mathcal{I}_P} \mathbf{\Lambda}_\alpha H_\alpha(\theta)$, we then have to solve a system of equations of size $M \times P$. It is the same system as (12) where we replace $\mathbf{A}$ by the reduced random matrix $\mathbf{W}^T\mathbf{A}\mathbf{W}$ and $\mathbf{b}$ by the reduced random vector $\mathbf{W}^T\mathbf{b}$. With an abuse of notation, we will denote by $\mathbf{\Lambda} = (\mathbf{W}^T\mathbf{A}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{b}$ the solution of this updating. Algorithm 2, named (PU-GSD), is a modified version of Algorithm 1 where we introduce the updating of stochastic functions after the construction of each couple $(\lambda_i, \mathbf{U}_i)$. With this algorithm, we try to compute the decomposition defined in definition 3.

Algorithm 2 *Power-type Generalized Spectral Decomposition with Updating (PU-GSD)*

1: $\mathbf{u} := 0$, $\tilde{\mathbf{b}} := \mathbf{b}$
2: ***for*** $i = 1$ *to* M ***do***
3: ***do*** *step 3 to 13 of algorithm 1 to compute* $\mathbf{U}_i$

4: $\mathbf{W} := (\mathbf{U}_1 \ ... \ \mathbf{U}_i)$
5: $\mathbf{\Lambda} := (\mathbf{W}^T\mathbf{A}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{b}$
6: $\mathbf{u} := \mathbf{W}\mathbf{\Lambda}$
7: $\tilde{\mathbf{b}} = \mathbf{b} - \mathbf{A}\mathbf{W}\mathbf{\Lambda}$
8: *Compute error indicator* $\epsilon^{(i)}$
9: ***if*** $\epsilon^{(i)} < \epsilon_{stop}$ ***then***
10: *break*
11: ***end if***
12: ***end for***

Regarding the results of proposition 4, we can here use an error criterium based on the evaluation of the generalized Rayleigh quotient

$$\epsilon_{ray}^{(i)} = \left| \frac{R(\mathbf{W}^{(i)}) - R(\mathbf{W}^{(i-1)})}{R(\mathbf{W}^{(i)})} \right|, \tag{47}$$

where $\mathbf{W}^{(j)} = (\mathbf{U}_1 \dots \mathbf{U}_j)$. We recall that at step 8, the generalized Rayleigh quotient can be simply computed as follows $R(\mathbf{W}) = Trace(E(\mathbf{\Lambda}\mathbf{b}^T\mathbf{W}))$.

6 Examples

The following three examples illustrate the efficiency of the proposed method on model problems. In example 1, the method is applied to a classical stationary heat diffusion problem with random source terms and a conductivity parameter which is modeled by a random field. In example 2, we consider the same problem but with a conductivity parameter modeled by a random variable. It is a degenerate case of the previous example (simple form of the random matrix), for which the solution is exactly represented with a decomposition of order 2. It illustrates that the proposed algorithms allow to automatically construct this exact decomposition. Finally, example 3 illustrates that the proposed algorithms are still efficient in the non-symmetric case.

6.1 Example 1

6.1.1 Description of the problem

[Fig. 1 about here.]

As a first model problem, we consider a classical stationary heat diffusion problem defined on a L-shaped spatial domain Ω (see figure 1). We denote by $u(\underline{x}, \theta)$ the temperature field. The normal flux g is imposed on a part $\partial_2\Omega$ of the boundary and the temperature is imposed to be zero on the complementary

part $\partial_1\Omega$. A volumic heat source f is also imposed on Ω. The space of admissible functions used in formulation (1) is $\mathcal{V} = \{v(\underline{x}) \in H^1(\Omega)\ ;\ v = 0\ on\ \partial_1\Omega\}$. In variational formulation (2), a and b forms are respectively

$$a(u, v; \theta) = \int_\Omega \kappa \underline{\nabla} u \cdot \underline{\nabla} v \, dx,$$
$$b(v; \theta) = \int_\Omega f v \, dx + \int_{\partial_2\Omega} g v \, ds,$$

where $\kappa(\underline{x}, \theta)$ is the conductivity parameter. At the space level, we use a classical finite element approximation. The mesh is composed by 1200 four-nodes elements and 1281 nodes. Random matrix $\mathbf{A}$ and random vector $\mathbf{b}$, respectively defined in (7) and (8), have the following components:

$$(\mathbf{A})_{ij} = \int_\Omega \kappa \underline{\nabla} \varphi_i \cdot \underline{\nabla} \varphi_j \, dx$$
$$(\mathbf{b})_j = \int_\Omega f \varphi_j \, dx + \int_{\partial_2\Omega} g \varphi_j \, ds,$$

where the φ_i are the basis functions of the approximation space $\mathcal{V}_n \subset \mathcal{V}$.

6.1.2 Stochastic modeling

The conductivity parameter is modeled by a random field. The following definition is taken from [17], where it was used to illustrate a method for identification of non-Gaussian random fields. Here, this definition allows us to impose a given marginal distribution which simply ensures the ellipticity of the bilinear form. We take

$$\kappa(\underline{x}, \theta) = F^{-1}_{\Gamma_\delta} \circ \Phi(\gamma(\underline{x}, \theta))),$$

where γ is a normalized Gaussian second-order random field such that $E(\gamma(\underline{x}, \cdot)) = 0$ and $E(\gamma(\underline{x}, \cdot)^2) = 1$. Function $y \to \Phi(y)$ is the cumulative distribution function of a normalized Gaussian random variable and function $y \to F_{\Gamma_\delta}(y)$ is the cumulative distribution function of a Gamma random variable:

$$F_{\Gamma_\delta}(y) = \int_0^y \frac{\delta}{\Gamma(\delta)} (\delta t)^{\delta-1} e^{-\delta t} \, dt,$$
$$with \quad \Gamma(\delta) = \int_0^\infty t^{\delta-1} e^{-t} \, dt.$$

With this definition, κ has a Gamma marginal distribution with unitary mean and standard deviation $\sigma = 1/\sqrt{\delta}$. We take $\delta = 16$ so that $\sigma = 0.25$. The Gaussian random field $\gamma(\underline{x}, \theta)$ is here defined by

$$\gamma(\underline{x}, \theta) = \sum_{k=1}^{3} \sqrt{\eta_k} \xi_k(\theta) \tilde{V}_k(\underline{x}), \tag{48}$$
$$with \quad \tilde{V}_k(\underline{x}) = \frac{V_k(\underline{x})}{\sqrt{\sum_{j=1}^3 \eta_j V_j(\underline{x})^2}},$$

where the $\xi_k \in \mathcal{N}(0,1)$ [3] are independent normalized Gaussian random variables and where $(\eta_k, V_k(\underline{x}))$ are the eigenpairs of the following homogeneous Fredholm equation of the second kind:

$$\int_\Omega exp(-\frac{\|\underline{x}-\underline{y}\|^2}{L^2})V_k(\underline{y})dy = \eta_k V_k(\underline{x}). \tag{49}$$

Then, the random field γ corresponds to a rescaled truncated Karhunen-Loève expansion of a Gaussian random field with exponential square correlation function. We take $L = 0.5$. In practice, problem (49) can be approximated by using finite elements. Here, we use the same approximation as for the solution u. We then have to solve a classical algebraic eigenvalue problem (cf. [5]). We use a classical technique to find its 3 rightmost eigenpairs.

Volumic heat source f and normal flux g are taken independent of the variable $\underline{x}$: $f(\underline{x},\theta) = \xi_4(\theta) \in \mathcal{N}(0.5, 0.2)$ and $g(\underline{x},\theta) = \xi_5(\theta) \in \mathcal{N}(0, 0.2)$, where ξ_4 and ξ_5 are independent Gaussian random variables, also independent of $\{\xi_i\}_{i=1}^3$.
The source of randomness is then represented by $m = 5$ independent Gaussian random variables. We use a polynomial chaos approximation of degree $p = 6$ at the stochastic level. The dimension of approximation space $\mathcal{S}_P$ is then $P = \frac{(m+p)!}{m!p!} = 462$.
The random field κ is projected on $\mathcal{S}_P$: $\kappa(\underline{x},\theta) = \sum_{\alpha\in\mathcal{I}_P} \kappa_\alpha(\underline{x})H_\alpha(\theta)$ where space functions $\kappa_\alpha = E(\kappa H_\alpha)$ are computed using Gauss-Hermite quadrature for the integration at the stochastic level. Matrix $\mathbf{A}$ can then be written $\mathbf{A} = \sum_{\alpha\in\mathcal{I}_P} \mathbf{A}_\alpha H_\alpha$ with $(\mathbf{A}_\alpha)_{ij} = \int_\Omega \kappa_\alpha \underline{\nabla}\varphi_i \cdot \underline{\nabla}\varphi_j \, dx$.

6.1.3 Reference solution and error criteria

We denote by $\mathbf{u}^{(M)}$ the approximation of order M (14) obtained by the generalized spectral decomposition algorithms, namely the power-type algorithms (P-GSD) or (PU-GSD). We denote by $\mathbf{u}$ the reference solution, which is the solution of (11). To compute $\mathbf{u}$, system (12) is solved with a preconditioned conjugate gradient (PCG) (see section 3.2), with a stopping tolerance of 10^{-8}.
We introduce the following error indicators to evaluate the quality of the approximation:

$$\epsilon_{sol}^{(M)} = \frac{\|\mathbf{u}-\mathbf{u}^{(M)}\|}{\|\mathbf{u}\|}, \quad \epsilon_{sol,\mathbf{A}}^{(M)} = \frac{\|\mathbf{u}-\mathbf{u}^{(M)}\|_{\mathbf{A}}}{\|\mathbf{u}\|_{\mathbf{A}}}. \tag{50}$$

where $\|\cdot\|$ is the classical L^2-norm (9) and $\|\cdot\|_{\mathbf{A}}$ is the $\mathbf{A}$-norm defined in (34). The generalized spectral decomposition obtained by power-type algorithms will be also compared with the direct Karhunen-Loève spectral decomposition of the reference solution $\mathbf{u}$, denoted by (Direct-SD).

[3] $\mathcal{N}(\mu,\sigma)$ denotes the set of Gaussian random variables with mean μ and standard deviation σ

Remark 11 *Of course, approximations are first introduced by the spatial and stochastic discretizations. The error introduced by the stochastic discretization can be defined by the difference between our reference solution* $\mathbf{u}$ *and the solution of the semi-discretized problem* (6)*. For this example, we estimated this error by computing the discretized solution associated with a polynomial chaos of degree* $p = 8$*. The estimation of the relative error in* L^2*-norm is* $1.04\,10^{-2}$*. In this article, we don't focus on those errors and consider the reference solution as the solution of the fully discretized problem* (11)*. However, this remark indicates that a very small tolerance for the resolution of the discretized problem* (11) *is generally useless for engineering applications, for which the discretization error is often greater than* 1%*.*

6.1.4 Convergence of power-type algorithms

(P-GSD) and (PU-GSD) have two parameters which are associated with the iterative search of each couple of functions (steps 3 to 13 of Algorithm 1): γ_{stop}, which defines the stopping criterium, and k_{max}, which defines the maximum number of iterations. Figures 2(a) and 2(b) show the error $\epsilon_{sol}^{(M)}$ with respect to the order M of the decomposition for different values of these parameters. On one hand, figure 2(a) shows that a relatively coarse convergence criterium ($\gamma_{stop} \approx 0.1$) do not affect the quality of the decomposition. On another hand, figure 2(b) shows that for a fixed stopping criterium $\gamma_{stop} = 0.05$, a few iterations are sufficient, which is in fact related to the fast convergence of this iterative procedure. For the following tests, we will choose $\gamma_{stop} = 0.05$ and $k_{max} = 3$.

[Fig. 2 about here.]

Figures 3(a) and 3(b) compare the convergence of (P-GSD), (PU-GSD) and (Direct-SD) with respect to the order of expansion M. We can observe that (PU-GSD) has almost the same convergence rate as (Direct-SD). This figure also shows the importance of the updating of stochastic functions. Indeed, without knowing the solution *a priori*, algorithm (PU-GSD) leads to a spectral decomposition of the solution which has quite the same quality as a direct Karhunen-Loève decomposition of the reference solution. We also verify, as mentioned in section 4.3.4, that (Direct-SD), when compared with (PU-GSD), gives a better approximation with respect to the L^2-norm but a coarser approximation with respect to the $\mathbf{A}$-norm. We can also notice than with a decomposition of order $M = 4$, the error is less than 1%.

[Fig. 3 about here.]

Figures 4(a),4(b) and 4(c) show the marginal probability density functions (PDFs) of the approximation obtained by (PU-GSD) for different orders of expansion M. The three sub-figures correspond respectively to points P_1, P_2

and P_3 (see figure 1). We can observe that with an approximation of order only $M = 4$, the approximate PDFs fit very well the reference solution.

[Fig. 4 about here.]

Figures 5(a) and 5(b) compare error indicators ϵ_{ray} and ϵ_{res} which can be used to evaluate the convergence of algorithms (P-GSD) and (PU-GSD). ϵ_{res} is the norm of the residual, defined in (44), and ϵ_{ray} is the indicator based on the Rayleigh quotient, defined in (45) for (P-GSD) and in (47) for (PU-GSD). These error indicators are compared to error indicators ϵ_{sol} and $\epsilon_{sol,\mathbf{A}}$, defined in (50). We can see that all these indicators are equivalent. The great advantage of estimator ϵ_{ray} is that it leads to very low computational costs.

[Fig. 5 about here.]

6.1.5 Analysis of the generalized spectral decomposition

Figures 6(a) and 6(b) show the first 8 deterministic vectors obtained respectively by (P-GSD) and (PU-GSD). We observe that these algorithms yield to the construction of quite relevent deterministic vectors. Vectors 1 and 2 take respectively into account the volumic load f and the surface load g. Subsequent vectors seem to take into account the fluctuations of the conductivity parameter. The superiority of (PU-GSD) appears clearly on this figure. Indeed, vectors 5 and 7 obtained by (P-GSD) are very similar to vectors 1 and 2. In fact, they can be interpreted as correction terms for the first two modes. Algorithm (PU-GSD) seems to capture these modes with the first two vectors and does not require any further correction.

[Fig. 6 about here.]

Figure 7 shows the 8 first vectors computed by a direct Karhunen-Loève decomposition of the reference solution (Direct-SD). If we compare this figure with figure 6 (b), we can see that (PU-GSD) and (Direct-SD) lead to very similar decompositions. In fact, we can say that the proposed algorithm (PU-GSD) allows to obtain a spectral decomposition of the solution, which is very similar to the Karhunen-Loève expansion, without knowing the solution *a priori*.

[Fig. 7 about here.]

6.1.6 Calculation time and memory requirements

We now look at the computation time of the proposed algorithms. Algorithm (PCG) took $156\,s$ to compute the reference solution. Figure 8 shows the error with respect to the calculation time for (P-GSD), (PU-GSD) and (PCG)

algorithms. The convergence curves for (P-GSD) and (PU-GSD) correspond to a generalized spectral decomposition up to order $M = 20$. We observe that power-type algorithms lead to the same computational cost. (PU-GSD) converges faster with respect to the order M but the cost of one iteration is greater than for (P-GSD), due to the updating of stochastic functions. On this example, we can then conclude that (PU-GSD) is more efficient since it leads to a decomposition with a lower order M for the same accuracy and calculation time. We see that power-type algorithms are significatively superior to the classical (PCG) algorithm. To compute a decomposition of order $M = 4$, which leads to a relatively good accuracy, it only takes $3.5\,s$ with (PU-GSD). The same accuracy is reached in $40\,s$ with (PCG). The calculation time is then divided by 11 on this simple example.

[Fig. 8 about here.]

To store the spectral decomposition of order M, we need to store $M \times (P + n)$ floating-point numbers. For (PCG) we need to store $n \times P$ floating-point numbers. For an approximation of order $M = 4$, memory requirements are divided by around 85. In fact, the gain is much greater since (PCG) algorithm generally requires reorthogonalization of the search directions. In fact, the storage of a Krylov subspace of dimension η requires to store $\eta \times n \times P$ floating-point numbers.

6.2 Example 2

We consider the same problem as in example 1. The linear form $b(v;\theta)$ is unchanged. The sources f and g are still defined by $f(\underline{x},\theta) = \xi_4(\theta) \in \mathcal{N}(0.5, 0.2)$ and $g(\underline{x},\theta) = \xi_5(\theta) \in \mathcal{N}(0, 0.2)$. But now, we suppose that the conductivity is a simple uniform random variable $\kappa(x,\theta) = \xi_1 \in U(0.7, 1.3)$ [4]. Random variables $\{\xi_1, \xi_4, \xi_5\}$ are considered independent. We use a generalized polynomial chaos approximation of degree $p = 8$ at the stochastic level [21]. We then use Legendre polynomials in the first stochastic dimension and Hermite polynomials in the two other stochastic dimensions. The dimension of approximation space $\mathcal{S}_P$ is then $P = \frac{(m+p)!}{m!p!} = 165$, where $m = 3$.

6.2.1 Convergence of power-type algorithms

Figure 9 shows the convergence with respect to the order of decomposition M for (P-GSD), (PU-GSD) and the Direct Karhunen-Loève decomposition of the solution (Direct-SD) . We observe that the three algorithms lead to the exact solution with a decomposition of order 2. That could have been expected

[4] $U(a,b)$ is the set of uniform random variables with value in $]a,b[$.

since the random matrix can be written $\mathbf{A} = \xi_1 \mathbf{A}_1$ where $\mathbf{A}_1$ is a deterministic matrix and the right hand side can be written $\mathbf{b} = \xi_4 \mathbf{b}_4 + \xi_5 \mathbf{b}_5$, where $\mathbf{b}_4$ and $\mathbf{b}_5$ are deterministic vectors.

[Fig. 9 about here.]

Figures 10(a) and 10(b) show the influence of parameters γ_{stop} and k_{max} of power-type algorithms. The iterative search of the couples $(\lambda_i, \mathbf{U}_i)$ converges very fast. We also observe that with a relatively coarse stagnation criterium ($\gamma_{stop} = 0.05$) and only two iterations ($k_{max} = 2$), we still obtain the exact numerical solution at order $M = 2$. Parameters γ_{stop} and k_{max} have a small influence on the obtained couples $(\mathbf{U}_1, \lambda_1)$ and $(\mathbf{U}_2, \lambda_2)$ but have no influence on the obtained decomposition of order $M = 2$. In fact, the iterative procedure converges in only 1 iteration for the construction of couple $(\mathbf{U}_2, \lambda_2)$.

[Fig. 10 about here.]

Here, both power-type algorithms required a calculation time of $0.7\,s$ while (PCG) required $4\,s$. The gain is still significant. However, comparing calculation time on this simple example is not so relevant.

6.2.2 Analysis of the generalized spectral decomposition

Figures 11(a), 11(b) and 11(c) show the two deterministic vectors which are obtained respectively by (P-GSD), (PU-GSD) and (Direct-SD). The obtained approximation of order 2 $\mathbf{u}^{(2)}$ is the same for the three algorithms. However, (P-GSD) and (PU-GSD) lead to deterministic vectors which are different from the ones obtained by (Direct-SD). It simply illustrates the fact that the decomposition is not unique. If we normalize all the deterministic vectors with the same norm and compare the vectors obtained by power-type algorithms and the ones obtained by (Direct-SD), we find a relative error of 2% for $\mathbf{U}_1$ and 5% for $\mathbf{U}_2$. The difference is due to the fact that the first ones are the optimal vectors with respect to the $\mathbf{A}$-norm and the second ones are the optimal vectors with respect to the L^2-norm.

[Fig. 11 about here.]

Stochastic functions λ_1 and λ_2 obtained by (P-GSD) and (PU-GSD) are also identical. Figures 12 and 13 show respectively the probability density functions of these two stochastic functions and their joint probability density function.

[Fig. 12 about here.]

[Fig. 13 about here.]

6.3 Example 3

In this last example, we briefly illustrate the fact that power-type algorithms also lead to satisfactory performances in the non-symmetric case. We here consider the same domain and boundary conditions as in example 1. The linear form $b(v;\theta)$ is unchanged but we take the following non symmetric bilinear form:

$$a(u,v;\theta) = \int_\Omega \kappa \underline{\nabla} u \cdot \underline{\nabla} v \, dx - \int_\Omega \underline{\chi}\, u \cdot \underline{\nabla} v \, dx,$$

where $\underline{\chi} = (\chi_1, \chi_2)^T$. We consider that material parameters are uniform random variables: $\kappa(\underline{x},\theta) = \xi_1(\theta) \in U(0.7, 1.3)$, $\chi_1(\underline{x},\theta) = \xi_2(\theta) \in U(5.5, 6.5)$, $\chi_2(\underline{x},\theta) = \xi_3(\theta) \in U(5.5, 6.5)$. The five random variables $\{\xi_i\}_{i=1}^5$ are independent. We use a generalized polynomial chaos approximation of degree $p = 5$. We then use Legendre polynomials in the first three stochastic dimensions and Hermite polynomials in the last two stochastic dimensions. We can notice that $\mathbf{A}$ still defines a norm $\| \ \|_{\mathbf{A}}$ since its symmetric part is almost surely positive definite. The reference solution, which solves system (12), is computed with a preconditionned conjugate gradient square algorithm (PCGS) with a stopping tolerance of 10^{-8}.

6.3.1 Convergence of power-type algorithms

We take $\gamma_{stop} = 0.05$ and $k_{max} = 4$ for the parameters of power-type algorithms. Figure 14(a) and 14(b) show the convergence of the spectral decomposition obtained by (P-GSD), (PU-GSD) and (Direct-SD). The three algorithms have quite the same convergence rate, which confirms the fact that power-type algorithms are also adapted to the non-symmetric case. The superiority of (PU-GSD) over (P-GSD) is not significant in this case.

[Fig. 14 about here.]

The resolution with (PCGS) took here $16\,s$ to reach a relative error of 10^{-8} and $2.8\,s$ to reach a relative error of 10^{-2}. The construction of a decomposition of order $M = 5$, which leads to a relative error lower than 10^{-2}, took $1.9\,s$ with (P-GSD) and $2.3\,s$ with (PU-GSD). For the same precision, the gain obtained in terms of computational time is not significant on this small-scale example. However, the gain in terms of memory requirements is still significant.

6.3.2 Analysis of the generalized spectral decomposition

Figures 15(a), 15(b) and 16 show the deterministic vectors obtained respectively by (P-GSD), (PU-GSD) and the direct Karhunen-Loève decomposition

of the solution (Direct-SD). We can observe that (P-GSD) and (PU-GSD) lead to quite the same vectors but than these vectors are significatively different from those obtained by (Direct-SD) (except for the first two vectors). However, we must recall than the obtained approximations $\mathbf{u}^{(M)}$ have quite the same accuracy although the deterministic vectors and stochastic functions are different.

[Fig. 15 about here.]

[Fig. 16 about here.]

7 Conclusion

We proposed a new method for solving stochastic partial differential equations based on a generalized spectral decomposition technique. The decomposition is the solution of a problem which can be interpreted as an "extended" generalized eigenproblem. This method allows to obtain a decomposition of the solution which is very similar to the one obtained by a classical truncated Karhunen-Loève expansion of the solution. Power-type algorithms have been proposed to solve this "extended" eigenproblem. The proposed resolution technique leads to significant computational savings compared to Krylov-type techniques which are classically used in the context of Galerkin Stochastic Finite Element methods. It also leads to a drastic reduction of memory requirements. It could then allows to deal with large-scale engineering problems and large stochastic dimensionality. Future works will be devoted to the development of more efficient algorithms to solve the "extended" eigenproblem, particularly for non-symmetric problems. We will also focus on the resolution of evolution equations and non-linear problems. This new method should be very efficient in these last cases since it enables the construction of a relevant reduced basis of deterministic functions which can be efficiently reused for subsequent resolutions.

Acknowledgement

This work is supported by the French National Research Agency (ANR). The author would like to thank Professor Pierre Ladevèze for fruitful discussions which helped to initiate the method proposed in this article.

References

[1] R. E. Caflisch, Monte carlo and quasi-monte carlo methods, Acta. Numer. 7 (1998) 1–49.

[2] M. Papadrakakis, V. Papadopoulos, Robust and efficient methods for stochastic finite element analysis using monte carlo simulation, Computer Methods in Applied Mechanics and Engineering 134 (1996) 325–340.

[3] B. Puig, F. Poirion, C. Soize, Non-gaussian simulation using hermite polynomial expansion: convergences, Probabilistic Engineering Mechanics 17 (2002) 253–264.

[4] M. Berveiller, B. Sudret, M. Lemaire, Stochastic finite element: a non intrusive approach by regression, European Journal of Computational Mechanics 15 (2006) 81–92.

[5] R. Ghanem, P. Spanos, Stochastic finite elements: a spectral approach, Springer, Berlin, 1991.

[6] M. Deb, I. Babuška, J. T. Oden, Solution of stochastic partial differential equations using galerkin finite element techniques, Computer Methods in Applied Mechanics and Engineering 190 (2001) 6359–6372.

[7] H. G. Matthies, A. Keese, Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations, Computer Methods in Applied Mechanics and Engineering 194 (12-16) (2005) 1295–1331.

[8] R. G. Ghanem, R. M. Kruger, Numerical solution of spectral stochastic finite element systems, Computer Methods in Applied Mechanics and Engineering 129 (1996) 289–303.

[9] M. F. Pellissetti, R. G. Ghanem, Iterative solution of systems of linear equations arising in the context of stochastic finite elements, Advances in Engineering Software 31 (2000) 607–616.

[10] A. Keese, H. G. Mathhies, Hierarchical parallelisation for the solution of stochastic finite element equations, Computer Methods in Applied Mechanics and Engineering 83 (2005) 1033–1047.

[11] P. Ladevèze, Nonlinear Computational Structural Mechanics - New Approaches and Non-Incremental Methods of Calculation, Springer Verlag, 1999.

[12] A. Nouy, P. Ladevèze, Multiscale computational strategy with time and space homogenization: a radial-type approximation technique for solving micro problems, International Journal for Multiscale Computational Engineering 170 (2) (2004) 557–574.

[13] A. Nouy, F. Schoefs, N. Moës, X-SFEM, a computational technique based on X-FEM to deal with random shapes, European Journal of Computational Mechanics 16 (2007) 277–293.

[14] I. Babuška, R. Tempone, G. E. Zouraris, Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation, Computer Methods in Applied Mechanics and Engineering 194 (2005) 1251–1294.

[15] P. Frauenfelder, C. Schwab, R. A. Todor, Finite elements for elliptic problems with stochastic coefficients, Computer Methods in Applied Mechanics and Engineering 194 (2-5) (2005) 205–228.

[16] A. D. R. Ghanem, On the construction and analysis of stochastic models: characterization and propagation of the errors associated with limited data, Journal of Computational Physics 217 (1) (2006) 63–81.

[17] C. Desceliers, R. Ghanem, C. Soize, Maximum likelihood estimation of stochastic chaos representations from experimental data, Int. J. for Numerical Methods in Engineering 66 (6) (2005) 978–1001.

[18] I. Babuška, P. Chatzipantelidis, On solving elliptic stochastic partial differential equations, Computer Methods in Applied Mechanics and Engineering 191 (2002) 4093–4122.

[19] B. Øksendal, Stochastic Differential Equations. An Introduction with Applications, fifth ed., Springer-Verlag, 1998.

[20] N. Wiener, The homogeneous chaos, Am. J. Math. 60 (1938) 897–936.

[21] D. B. Xiu, G. E. Karniadakis, The Wiener-Askey polynomial chaos for stochastic differential equations, SIAM J. Sci. Comput. 24 (2) (2002) 619–644.

[22] O. P. Le Maître, O. M. Knio, H. N. Najm, R. G. Ghanem, Uncertainty propagation using Wiener-Haar expansions, Journal of Computational Physics 197 (2004) 28–57.

[23] C. Soize, R. Ghanem, Physical systems with random uncertainties: chaos representations with arbitrary probability measure, SIAM J. Sci. Comput. 26 (2) (2004) 395–410.

[24] J. T. Oden, I. Babuska, F. Nobile, Y. Feng, R. Tempone, Theory and methodology for estimation and control of errors due to modeling, approximation, and uncertainty, Computer Methods in Applied Mechanics and Engineering 194 (2-5) (2005) 195–204.

[25] P. Ladevèze, E. Florentin, Verification of stochastic models in uncertain environments using the constitutive relation error method, Computer Methods in Applied Mechanics and Engineering 196 (1-3) (2005) 225–234.

[26] A. Sameh, Z. Tong, The trace minimization method for the symmetric generalized eigenvalue problem, J. Comput. Appl. Math. 123 (2000) 155–175.

List of Figures

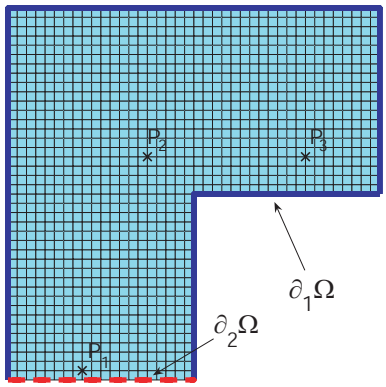
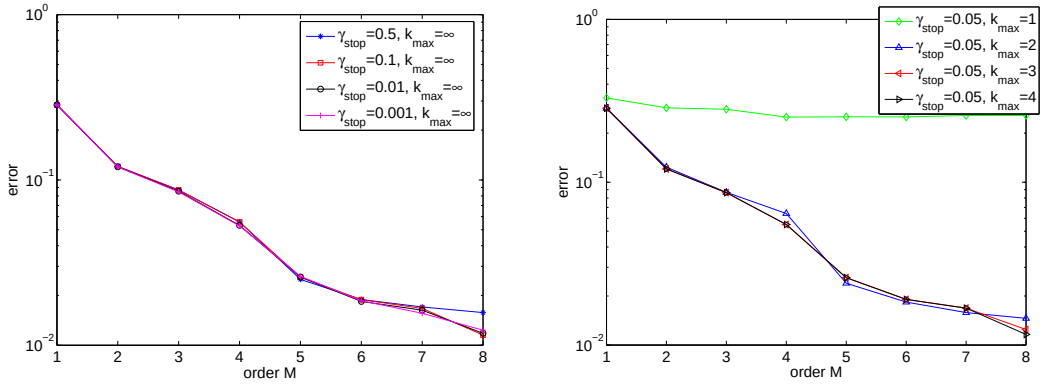
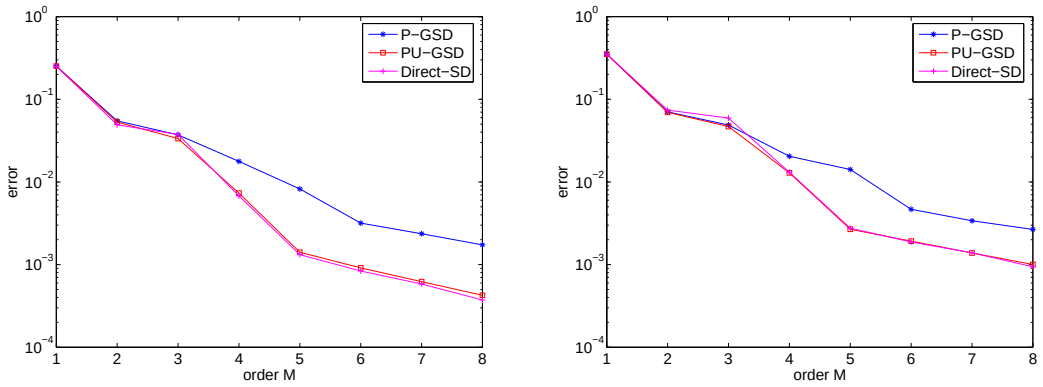


Fig. 1. Description of example 1: mesh and boundary conditions



(a) Influence of γ_{stop} (b) Influence of k_{max}

Fig. 2. Relative error with respect to the order of decomposition M for (P-GSD). Influence of parameters γ_{stop} (a) and k_{max} (b)



(a) ϵ_{sol} (b) $\epsilon_{sol,\mathbf{A}}$

Fig. 3. Relative error with respect to the order of decomposition M for (P-GSD), (PU-GSD) and (Direct-SD). (a) L^2-norm, (b) $\mathbf{A}$-norm

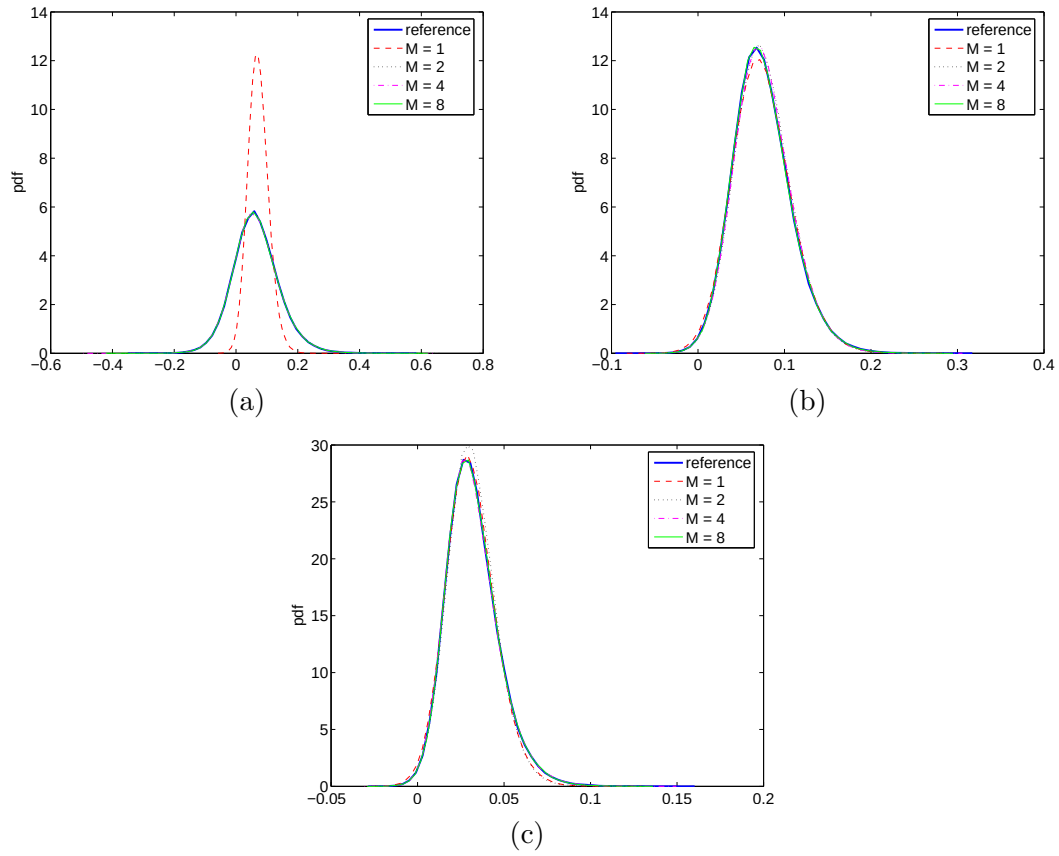
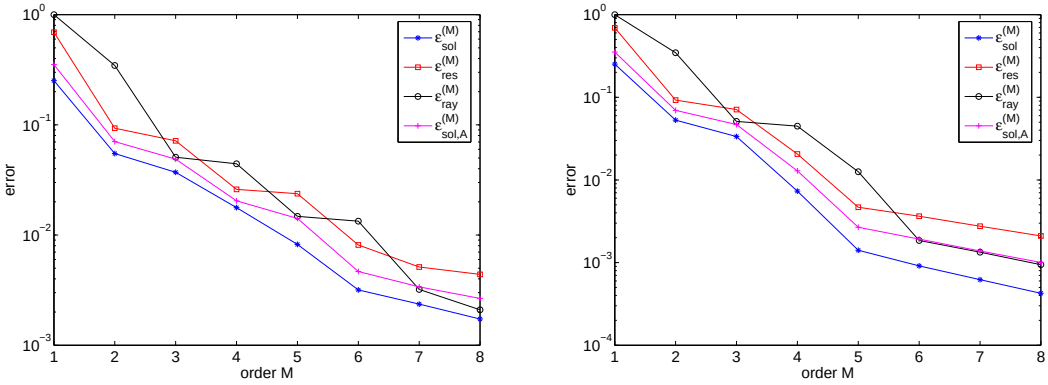
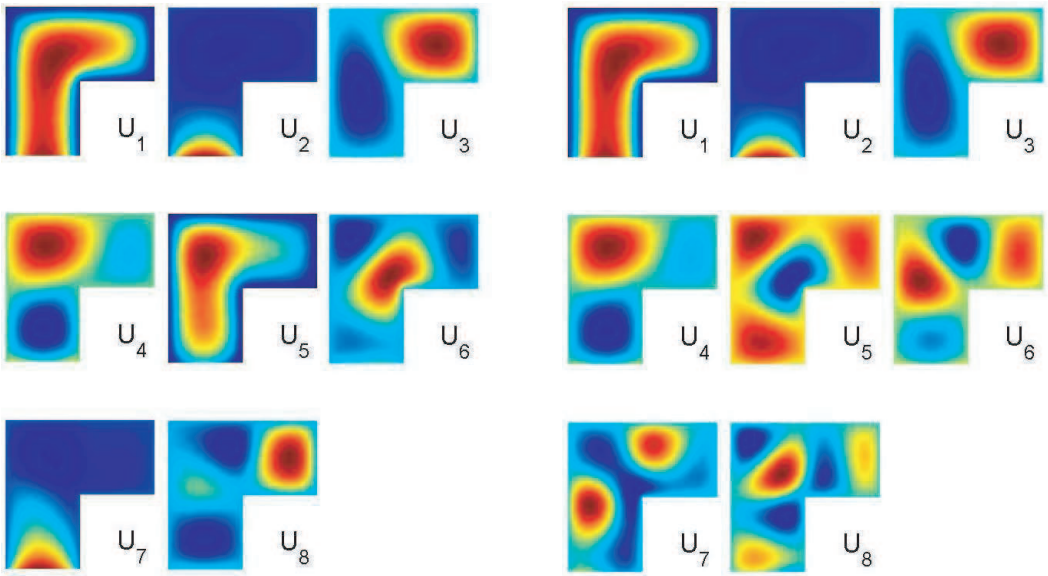


Fig. 4. (PU-GSD): marginal PDFs of u^M at points P_1 (a), P_2 (b) and P_3 (c)



(a) (P-GSD) (b) (PU-GSD)

Fig. 5. Error indicators for (P-GSD) (a) and (PU-GSD) (b)



(a) (P-GSD) (b) (PU-GSD)

Fig. 6. Vectors $\{\mathbf{U}_i\}_{i=1}^{8}$ obtained by (P-GSD) (a) and (PU-GSD) (b)

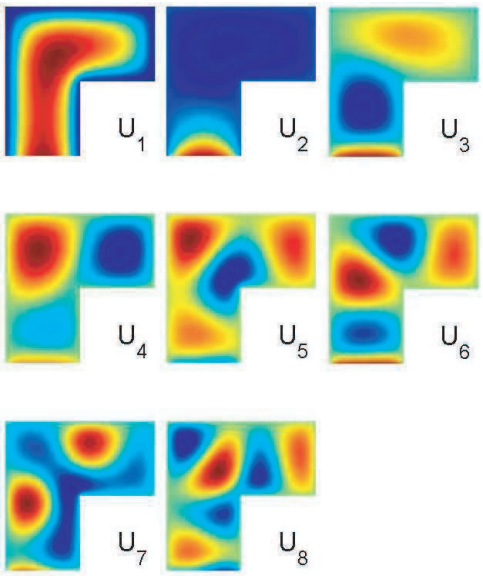


Fig. 7. Vectors $\{\mathbf{U}_i\}_{i=1}^{8}$ obtained by (Direct-SD)

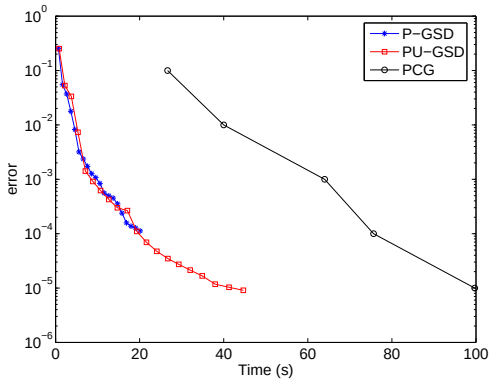


Fig. 8. Error ϵ_{sol} versus computation time for algorithms (P-GSD), (PU-GSD) and (PCG)

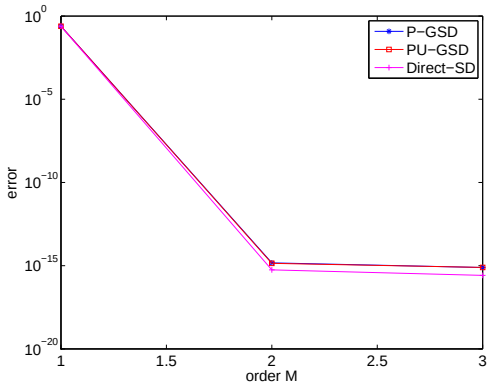
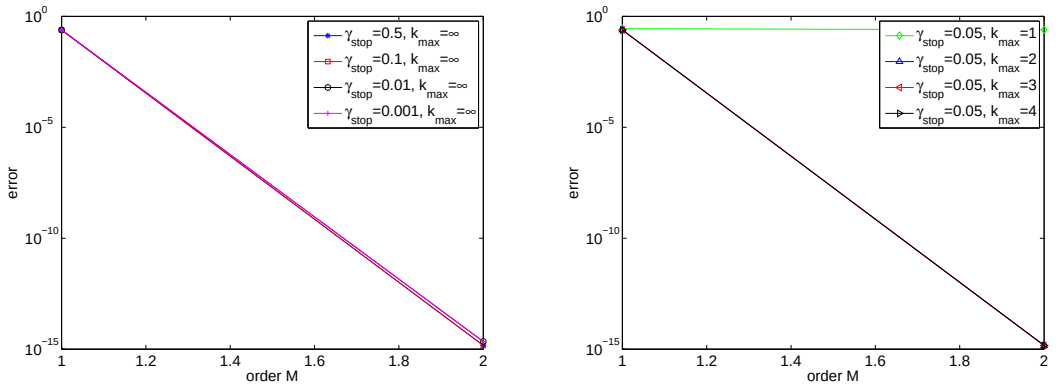
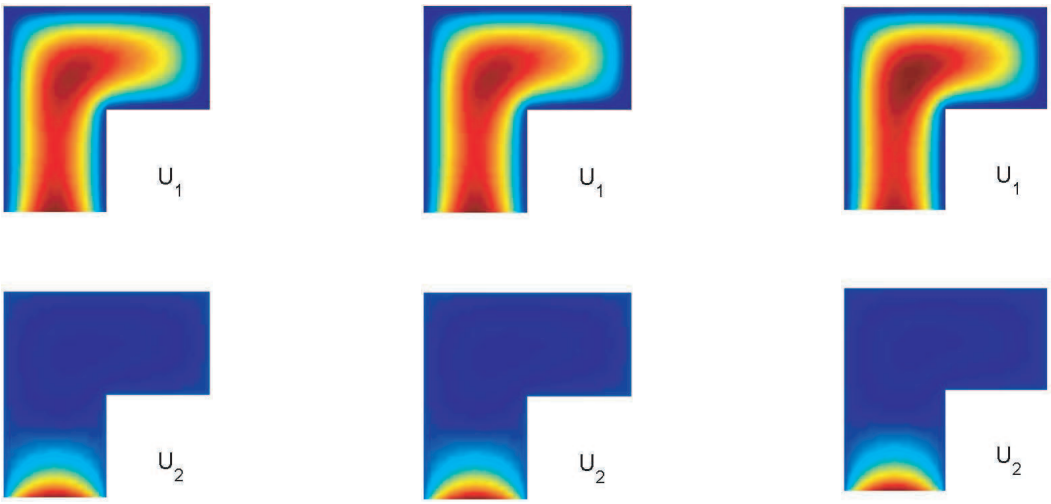


Fig. 9. Error in **A**-norm with respect to the order of decomposition M for (P-GSD), (PU-GSD) and (Direct-SD)



(a) Influence of γ_{stop} (b) Influence of k_{max}

Fig. 10. Relative error with respect to the order of decomposition M for (P-GSD). Influence of parameters γ_{stop} (a) and k_{max} (b)



(a) (P-GSD) (b) (PU-GSD) (c) (Direct-SD)

Fig. 11. Vectors $\mathbf{U}_1$ and $\mathbf{U}_2$ obtained by (P-GSD) (a), (PU-GSD) (b) and (Direct-SD) (c)

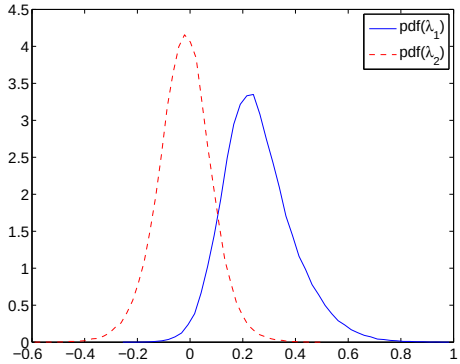


Fig. 12. Probability density functions of λ_1 and λ_2

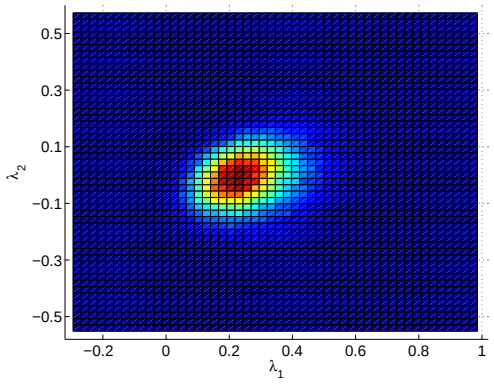
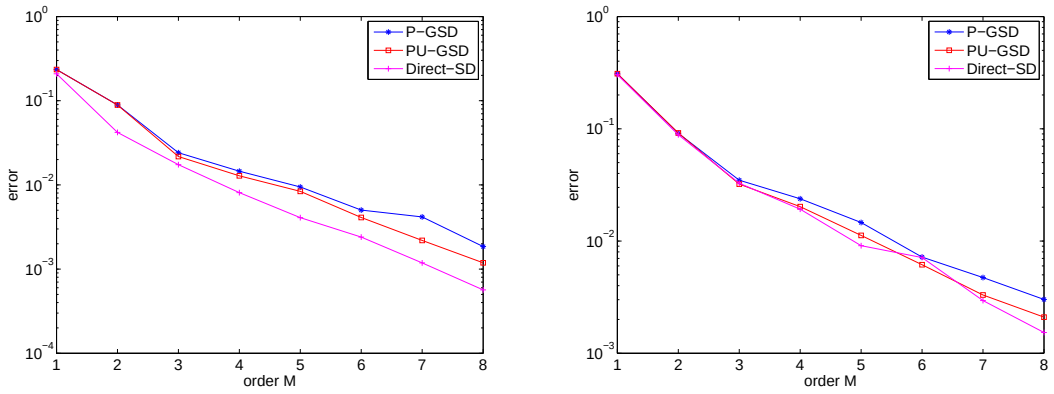
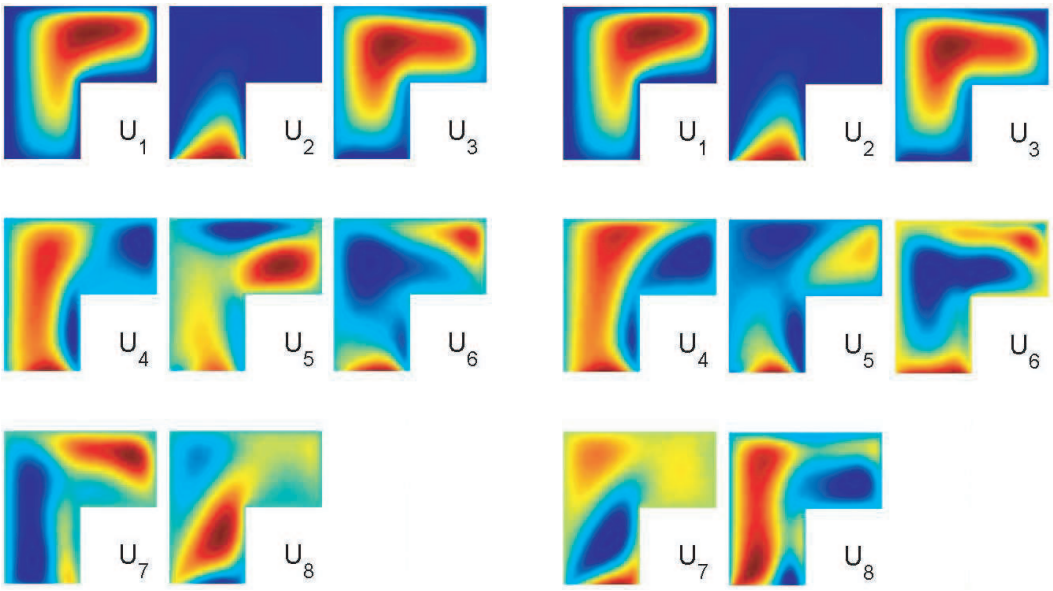


Fig. 13. Joint probability density function of λ_1 and λ_2



(a) ϵ_{sol} (b) $\epsilon_{sol,\mathbf{A}}$

Fig. 14. Relative error with respect to the order of decomposition M for (P-GSD), (PU-GSD) and (Direct-SD). (a) L^2-norm, (b) $\mathbf{A}$-norm



(a) (P-GSD) (b) (PU-GSD)

Fig. 15. Vectors $\{\mathbf{U}_i\}_{i=1}^{8}$ obtained by (P-GSD) (a) and (PU-GSD) (b)

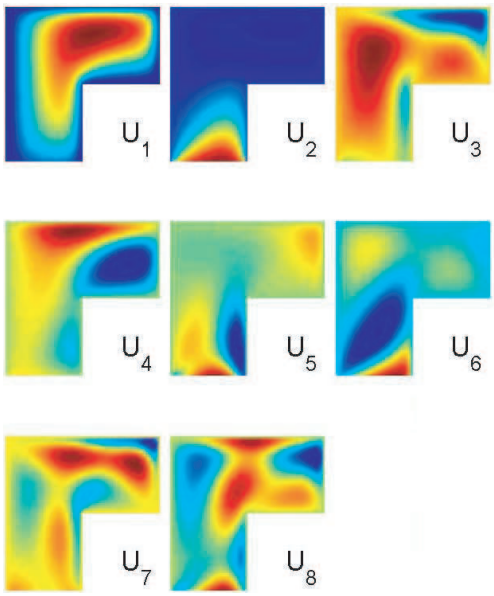


Fig. 16. Vectors $\{\mathbf{U}_i\}_{i=1}^{8}$ obtained by (Direct-SD)