Compressed principal component analysis of non-Gaussian vectors

Marc Mignolet, Christian Soize

COMPRESSED PRINCIPAL COMPONENT ANALYSIS OF NON-GAUSSIAN VECTORS

MARC MIGNOLET* AND CHRISTIAN SOIZE†

Abstract. A novel approximate representation of non-Gaussian random vectors is introduced and validated, which can be viewed as a Compressed Principal Component Analysis (CPCA). This representation relies on the eigenvectors of the covariance matrix obtained as in a Principal Component Analysis (PCA) but expresses the random vector as a linear combination of a random sample of N of these eigenvectors. In this model, the indices of these eigenvectors are independent discrete random variables with probabilities proportional to the corresponding eigenvalues. Moreover, the coefficients of the linear combination are zero mean unit variance random variables. Under these conditions, it is first shown that the covariance matrix of this CPCA matches exactly its PCA counterpart independently of the value of N. Next, it is also shown that the distribution of the random coefficients can be selected, without loss of generality, to be a symmetric function. Then, to represent the vector of these coefficients, a novel set of symmetric vector-valued multidimensional polynomials of the canonical Gaussian random vector is derived. Interestingly, it is noted that the number of such polynomials is only slowly growing with the maximum polynomial order thereby providing a framework for a compact approximation of the target random vector. The identification of the deterministic parameters of the expansion of the random coefficients on these symmetric vector-valued multidimensional polynomial is addressed next. Finally, an example of application is provided that demonstrates the good matching of the distributions of the elements of the target random vector and its approximation with only a very limited number of parameters.

Key words. Principal component analysis, Compressed principal component analysis, Non-Gaussian vector, Random eigenvectors, Symmetric polynomials, Random fields, Stochastic processes, Inverse problem, Stochastic model, Reduction method, Uncertainty quantification, Stochastic modeling

AMS subject classifications. 00A20

1. Introduction.

The objective of this paper is to propose the *Compressed Principal Component Analysis* (CPCA) that is a novel small parameterized representation of any non-Gaussian second-order random variable $\mathbf{X} = (X_1, \ldots, X_n)$ with values in $\mathbb{R}^n$. This representation would be useful for solving statistical inverse problems related to any stochastic computational model for which there is an uncertain vector-valued system-parameter that is modeled by a random vector $\mathbf{X}$.

To explain the benefits of this representation, consider the framework of a classical statistical inverse problem. Let us assume that a parameterized representation of $\mathbf{X}$ has been constructed and is written as $\mathbf{X} = \mathbf{g}(\mathbf{z}, \boldsymbol{\Xi})$ in which $\boldsymbol{\Xi} = (\Xi_1, \ldots, \Xi_N)$ is the $\mathbb{R}^N$-valued normalized Gaussian random variable (centered and with a covariance matrix that is the identity matrix) the probability distribution of which is denoted by $P_{\boldsymbol{\Xi}}(d\boldsymbol{\xi})$ on $\mathbb{R}^N$. The parameterization of the representation corresponds to the vector $\mathbf{z} = (z_1, \ldots, z_M)$ of hyperparameters, which belongs to an admissible set that is a subset $\mathcal{C}_{\mathbf{z}}$ of $\mathbb{R}^M$. The measurable mapping $\boldsymbol{\xi} \mapsto \mathbf{g}(\mathbf{z}, \boldsymbol{\xi})$ is defined through the construction of the representation. Consequently, if $\mathbf{z}$ is fixed to a given value $\mathbf{z}^{\text{opt}}$, then the probability distribution $P_{\mathbf{X}}$ of $\mathbf{X}$ is completely defined as the image of $P_{\boldsymbol{\Xi}}(d\boldsymbol{\xi})$ under the mapping $\boldsymbol{\xi} \mapsto \mathbf{g}(\mathbf{z}^{\text{opt}}, \boldsymbol{\xi})$. Let us consider a computational model with an uncertain system-parameter $\mathbf{x}$ that is modeled by random variable $\mathbf{X}$. Let $\mathbf{Q}$ be the vector-valued random quantity of interest that is constructed as an obser-

*Arizona State University, SEMTE, Faculties of Mechanical and Aerospace Engineering, Tempe, AZ 85287-6106, USA (marc.mignolet@asu.edu).

†Université Gustave Eiffel, Modélisation et Simulation Multi-Echelle, MSME UMR 8208, 5 bd Descartes, 77454 Marne-la-Vallée, France (christian.soize@univ-eiffel.fr).

vation of the random output of the stochastic computational model, and which can be written as $\mathbf{Q} = \mathbf{f}(\mathbf{X}) = \mathbf{f}(\mathbf{g}(\mathbf{z}, \boldsymbol{\Xi}))$. It is assumed that the measurable mappings $\mathbf{f}$ and $\mathbf{g}$ are such that the probability distribution $P_{\mathbf{Q}}(d\mathbf{q}; \mathbf{z})$ of $\mathbf{Q}$ given $\mathbf{z}$ in $\mathcal{C}_{\mathbf{Z}}$ admits a probability density function $p_{\mathbf{Q}}(\mathbf{q}; \mathbf{z})$ with respect to the Lebesgue measure $d\mathbf{q}$. Let a target be available for the QoI, which corresponds to n_r independent realizations $\{\mathbf{q}^{\exp,r}, r = 1, \ldots, n_r\}$ coming for instance from experiments or from simulations. Several statistical approaches exist for solving statistical inverse problems. We refer the reader to [24, 52, 17, 45] for an overview concerning the general methodologies for statistical and computational inverse problems, including general least-square inversion and the maximum likelihood method [40, 49], and including the Bayesian approach [49, 9, 7, 51]. To simplify the presentation, let us assume that the maximum likelihood method is used for identifying an optimal value $\mathbf{z}^{\text{opt}}$ in $\mathcal{C}_{\mathbf{z}}$ of hyperparameter $\mathbf{z}$,

$$\mathbf{z}^{\text{opt}} = \arg\max_{\mathbf{z} \in \mathcal{C}_{\mathbf{z}}} L(\mathbf{z}) \,, \tag{1.1}$$

in which $L(\mathbf{z}) = \sum_{r=1}^{n_r} \log p_{\mathbf{Q}}(\mathbf{q}^{\exp,r}, \mathbf{z})$ is the log-likelihood function defined on $\mathcal{C}_{\mathbf{z}}$. For any given $\mathbf{z}$ in $\mathcal{C}_{\mathbf{z}}$ and for r in $\{1, \ldots, n_r\}$, the value $\log p_{\mathbf{Q}}(\mathbf{q}^{\exp,r}, \mathbf{z})$ of the log-pdf is estimated with the computational model (or a surrogate model derived from the computational model) and the known canonical Gaussian density $p_{\boldsymbol{\Xi}}$ of $\boldsymbol{\Xi}$. This is accomplished using a stochastic solver, e.g., a Monte Carlo solver [39], and a density estimation technique such as the multivariate kernel density one [6, 22, 45]. In general, the optimization problem defined by Eq. (1.1) is not convex and consequently, the numerical cost for computing an approximation of $\mathbf{z}^{\text{opt}}$ increases with the dimension M. So, one challenge is to reduce the number M of coefficients that has to be identified using an adapted representation $\mathbf{X} = \mathbf{g}(\mathbf{z}, \boldsymbol{\Xi})$ and this reduction is the primary objective of this investigation.

The most general representation is the Polynomial Chaos Expansion (PCE) because it allows for representing any probability distribution $P_{\mathbf{X}}(d\mathbf{x})$ of any second-order non-Gaussian random field. In computational sciences and engineering, the development and the use of PCE for representing random fields have been pioneered by Roger Ghanem in 1990-1991 [20] who proposed to combine a Karhunen-Loeve expansion (that allows using a statistical reduced model) with a PCE of the statistical reduced model. This type of construction has then been re-analyzed and used for solving boundary value problems using the spectral approach (see for instance [37, 12, 15, 19, 21, 33, 25]). The PCE has also been extended for an arbitrary probability measure [59, 26, 27, 47, 58, 16]) and for sparse representation [3, 4, 5, 1, 30]. Further, new algorithms have been proposed for obtaining a robust computation of realizations of high degrees polynomial chaos [46, 35]. This type of representation has also been extended for the case of the polynomial chaos expansion with random coefficients [48], for the construction of a basis adaptation in homogeneous chaos spaces [55, 54, 57], and for an arbitrary multimodal multidimensional probability distribution [44]. It should be noted that the space sampling of a random field or the time sampling of a stochastic process yields a random vector for which all the above methods for the PCE can be used, the Karhunen-Loeve expansion being replaced by a Principal Component Analysis (PCA).

The use of the PCE for constructing a parameterized representation of a non-Gaussian random field that models the parameter of a boundary value problem, in order to identify it solving a statistical inverse problem has been initiated in [13, 18, 14] and used and revisited in [11, 23, 10, 42, 35, 36, 8, 34, 45] for statistical inverse problems in low or in high stochastic dimension. For the statistical inverse

identification of the coefficients of the PCE, the Bayesian approach has been proposed in [18, 32, 2, 43, 38, 29, 50, 41, 56].

Using the PCA, a classical finite approximation $\mathbf{X}^m_{\text{PCA}}$ of $\mathbf{X}$ yields the following representation $\mathbf{X}^m_{\text{PCA}} = \mathbf{m}_{\mathbf{X}} + \sum_{\alpha=1}^m \sqrt{\lambda_\alpha}\, \Gamma_\alpha\, \boldsymbol{\varphi}^\alpha$, in which $\mathbf{m}_{\mathbf{X}}$ is the mean vector of $\mathbf{X}$ and where λ_α and $\boldsymbol{\varphi}^\alpha$ are the eigenvalues and the eigenvectors of the covariance matrix of $\mathbf{X}$. The centered random vector $\boldsymbol{\Gamma} = (\Gamma_1, \ldots, \Gamma_m)$ whose covariance matrix is the identity matrix, can be represented by the truncated PCE, $\Gamma^{(M)}_\alpha = \sum_{k=1}^M z^k_\alpha\, \Psi_k(\boldsymbol{\Xi})$ in which $\{\Psi_k\}_k$ are the multivariate normalized Hermite polynomials for which the multi-indices are renumbered with single index k. In such a case, $\mathbf{X}^m_{\text{PCA}}$ is rewritten as $\mathbf{X}^{(m,M)}_{\text{PCA}}$. For $m = n$, the mean-square convergence of the sequence of random variables $\{\mathbf{X}^{(n,M)}_{\text{PCA}}\}_M$ is guaranteed as M goes to infinity. For a fixed value of m and M, the number of coefficients z^k_α that have to be identified is $m \times M$ and that can be exceedingly large for solving the optimization problem defined by Eq. (1.1). This is the reason why there is an interest for constructing a reduced representation, the CPCA, in order to minimize the number of coefficients that have to be identified in the representation.

Such dimensionality reduction of the representation, which allows for facilitating the identification of the coefficients, remains a challenge. To the author's knowledge, in addition to the sparse representation, two other approaches have been proposed in [31, 55]. In [55], the authors propose to adapt the PCE to the QoI by introducing a rotation of the Gaussian subspace. This rotation, which depends on the QoI, has to be identified for each QoI and provides a reduction of full polynomial expansions through the identified rotation. In this work, we propose an alternative that allows for drastically reducing the number of polynomials, and therefore for reducing the number of coefficients that have to be identified, independently of the QoI. This means that the reduced representation is constructed independently of the QoI and can then be used for any QoI. The main idea of this novel representation is to use symmetric polynomials associated with the canonical Gaussian measure, the number of which is very small with respect to all the Hermite polynomials. However, it can be seen that such symmetric polynomials cannot be used for representing the non-Gaussian random vector $\boldsymbol{\Gamma}$ that is involved in the PCA of $\mathbf{X}$ because they do not exhibit this symmetry property. Consequently, a novel PCA with random eigenvectors, referred to as CPCA, is proposed and is written as $\mathbf{X}^N_{\text{CPCA}} = \mathbf{m}_{\mathbf{X}} + \sum_{\alpha=1}^N \sqrt{\mu/N}\, H_\alpha\, \boldsymbol{\varphi}^{J_\alpha}$, in which $\mu = \sum_{\alpha=1}^m \lambda_\alpha$, where $J_1, \ldots, J_N$ are N independent copies of a random variable J with values in the set $\{1, \ldots, m\}$ of integers, and for which the probability distribution is proportional to eigenvalues $\lambda_1, \ldots, \lambda_m$. This CPCA will be detailed in Section 2 and is based on the use of a novel set $\{\boldsymbol{\psi}^k(\boldsymbol{\Xi})\}_k$ of symmetric vector-valued multidimensional polynomial of the canonical Gaussian random vector $\boldsymbol{\Xi}$. Accordingly, the random vector $\mathbf{H} = (H_1, \ldots, H_N)$ is written as $\mathbf{H} = \sum_{k=1}^M z_k\, \boldsymbol{\psi}^k(\boldsymbol{\Xi})$ in which the coefficients of the representation are gathered in vector $\mathbf{z} = (z_1, \ldots, z_M)$ whose length M is very small. This construction is presented in Section 3.

2. Compressed principal component analysis with random eigenvectors. Prior to introducing the novel compressed principal component analysis (CPCA) with random eigenvectors, it is useful to briefly review the standard Principal Component Analysis (PCA).

2.1. Principle used for the construction. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a second-order random vector defined on a probability space $(\Theta, \mathcal{T}, \mathcal{P})$ with values in $\mathbb{R}^n$ the probability distribution of which is $P_{\mathbf{X}}(d\mathbf{x}) = p_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$ where $\mathbf{x} \mapsto p_{\mathbf{X}}(\mathbf{x})$ is the prob-

ability density function defined on $\mathbb{R}^n$. Denoting by $E\{\dot{\}}$ the operator of mathematical expectation, the mean vector in $\mathbb{R}^n$ of $\mathbf{X}$ is $\mathbf{m}_{\mathbf{X}} = E\{\mathbf{X}\}$. Moreover, the covariance matrix of $\mathbf{X}$ is $[C_{\mathbf{X}}] = E\{(\mathbf{X} - \mathbf{m}_{\mathbf{X}})\,(\mathbf{X} - \mathbf{m}_{\mathbf{X}})^T\}$ where $.^T$ is the operation of matrix transposition.

The principal component representation $\mathbf{X}^m_{\text{PCA}}$ of the vector $\mathbf{X}$ is then

$$\mathbf{X}^m_{\text{PCA}} = \mathbf{m}_{\mathbf{X}} + \sum_{\alpha=1}^{m} \sqrt{\lambda_\alpha}\, \Gamma_\alpha\, \boldsymbol{\varphi}^\alpha\,, \tag{2.1}$$

where $m \leq n$, and for $\alpha = 1, \ldots, n$, λ_α and $\boldsymbol{\varphi}^\alpha = (\varphi^\alpha_1, \ldots, \varphi^\alpha_n)$ are the eigenvalues and eigenvectors of $[C_{\mathbf{X}}]$. Owing to the symmetry of the covariance matrix, the eigenvectors $\boldsymbol{\varphi}^1, \ldots, \boldsymbol{\varphi}^n$ form an orthonormal basis of $\mathbb{R}^n$. In addition, Γ_α are zero mean, unit variance uncorrelated random variables defined as

$$\Gamma_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} (\boldsymbol{\varphi}^\alpha)^T\, (\mathbf{X} - \mathbf{m}_{\mathbf{X}})\,. \tag{2.2}$$

The random vector $\mathbf{X}^m_{\text{PCA}}$ of Eq. (2.1) converges in mean square to $\mathbf{X}$ as $m \to n$ with error

$$\varepsilon_{\text{PCA}}(m) = \frac{E\{\|\mathbf{X} - \mathbf{X}^m_{\text{PCA}}\|^2\}}{E\{\|\mathbf{X}\|^2\}} = 1 - \frac{\sum_{\alpha=1}^m \lambda_\alpha}{\text{tr}[C_{\mathbf{X}}]} \tag{2.3}$$

where $\|.\|$ and tr[.] are the Euclidean norm of a vector and the trace of a matrix.

2.2. Definition of the CPCA with random eigenvectors. With the above notations, the compressed principal component approximation $\mathbf{X}^N_{\text{CPCA}}$ of the vector $\mathbf{X}$ is defined as

$$\mathbf{X}^N_{\text{CPCA}} = \mathbf{m}_{\mathbf{X}} + \sqrt{\frac{\mu}{N}} \sum_{\alpha=1}^{N} H_\alpha\, \boldsymbol{\varphi}^{J_\alpha}, \tag{2.4}$$

where N is a parameter of the approximation, which can be less than, equal to, or greater than m

$$\mu = \sum_{\alpha=1}^{m} \lambda_\alpha\ , \tag{2.5}$$

and the real-valued random variables $\{H\}_\alpha$ have zero mean, unit variance and are uncorrelated. Finally, $\mathbf{J} = (J_1, \ldots, J_N)$ denotes a random vector that is independent of $\mathbf{H}$ and for which the components $J_1, \ldots, J_N$ are N independent copies of a random variable J with values in the set $\{1, \ldots, m\}$ of integers and with distribution

$$\text{Prob}\{\text{J} = \alpha\} = \frac{\lambda_\alpha}{\mu} \quad \text{for} \quad \alpha = 1, \ldots, \text{m}\,. \tag{2.6}$$

2.3. Some observations and properties.

2.3.1 The random vector of Eqs (2.4) and (2.5) represents in general, e.g., for $m > 2N$, an approximation of the PCA random vector $\mathbf{X}$. The accuracy of this approximation will be quantified here by the overlap defined as

$$\epsilon_{ovl} = \frac{1}{n} \sqrt{\sum_{i=1}^{n} (\epsilon^{(i)}_{ovl})^2}\,, \tag{2.7}$$

with

$$\epsilon_{ovl}^{(i)} = \int_{-\infty}^{\infty} |p_{X_i}(x) - \widehat{p}_{X_i}(x)|\, dx \quad , \quad i = 1, \ldots, n\,, \tag{2.8}$$

where $p_{X_i}(x)$ is the probability density function of the component X_i corresponding to the model of Eqs (2.4) and $\widehat{p}_{X_i}(x)$ is the estimate of this function obtained from available data, simulated or experimental.

2.3.2 Combining Eqs (2.2) and (2.4), the random variable Γ_α of the PCA of $\mathbf{X}_{\text{CPCA}}^N$ is written as

$$\Gamma_\alpha = \sqrt{\frac{\mu}{N\lambda_\alpha}} \sum_{\beta=1}^{N} H_\beta\, (\boldsymbol{\varphi}^\alpha)^T \boldsymbol{\varphi}^{J_\beta} = \sqrt{\frac{\mu}{N\lambda_\alpha}} \sum_{\beta=1}^{N} H_\beta\, \delta_{\alpha J_\beta}\,, \tag{2.9}$$

where $\delta_{\alpha\beta}$ denotes the Kronecker symbol.

2.3.3 Given the distribution of the random indices J_α, Eq. (2.6), one has

$$E\{g(J_1, \ldots, J_p)\} = \sum_{\alpha_1=1}^{m} \ldots \sum_{\alpha_p=1}^{m} \frac{\lambda_{\alpha_1} \ldots \lambda_{\alpha_p}}{\mu^p}\, g(\alpha_1, \ldots, \alpha_p)\,. \tag{2.10}$$

Some noteworthy applications of the above property are as follows

$$E\{(\boldsymbol{\varphi}^{J_\alpha})^T \boldsymbol{\varphi}^{J_\beta}\} = \sum_{\alpha=1}^{m} \sum_{\beta=1}^{m} \frac{\lambda_\alpha \lambda_\beta}{\mu^2}\, (\boldsymbol{\varphi}^\alpha)^T \boldsymbol{\varphi}^\beta = \sum_{\alpha=1}^{m} \frac{\lambda_\alpha^2}{\mu^2}\,, \tag{2.11}$$

given the orthogonality of the eigenvectors $\boldsymbol{\varphi}^\alpha$ and when $\alpha \neq \beta$. When $\alpha = \beta$, the normality of the eigenvectors implies that $E\{(\boldsymbol{\varphi}^{J_\alpha})^T \boldsymbol{\varphi}^{J_\alpha}\} = 1$. Moreover, one also has from Eq. (2.10) that

$$E\{\boldsymbol{\varphi}^{J_\alpha} (\boldsymbol{\varphi}^{J_\alpha})^T\} = \sum_{\alpha=1}^{m} \frac{\lambda_\alpha}{\mu}\, \boldsymbol{\varphi}^\alpha (\boldsymbol{\varphi}^\alpha)^T. \tag{2.12}$$

Note in Eqs (2.11) and (2.12) that the left hand sides of these equations do not depend on α which is thus a dummy index.

Consider next the characteristic function of $\mathbf{X}_{\text{CPCA}}^N$ denoted as $\psi_N(\mathbf{u}) = E\{\exp(i\, \mathbf{u}^T \mathbf{X}_{\text{CPCA}}^N)\}$. Using Eq. (2.4) and splitting the expectation in expectations over the random variables H_α and J_α separately, denoted by $E_{\mathbf{H}}$ and $E_{\mathbf{J}}$, respectively, it is found that

$$\psi_N(\mathbf{u}) = \exp(i\, \mathbf{u}^T \mathbf{m}_{\mathbf{X}})\, E_{\mathbf{H}}\{E_{\mathbf{J}}\{\exp(i \sqrt{\frac{\mu}{N}} \sum_{\alpha=1}^{N} H_\alpha\, \mathbf{u}^T \boldsymbol{\varphi}^{J_\alpha})\}\}. \tag{2.13}$$

Moreover, using Eq. (2.10), the above equation can be rewritten as

$$\psi_N(\mathbf{u}) = \exp(i\, \mathbf{u}^T \mathbf{m}_{\mathbf{X}}) \sum_{\alpha_1=1}^{m} \ldots \sum_{\alpha_N=1}^{m} \frac{\lambda_{\alpha_1} \ldots \lambda_{\alpha_N}}{\mu^N}\, E_{\mathbf{H}}\{\exp(i \sqrt{\frac{\mu}{N}} \sum_{\beta=1}^{N} H_\beta\, \mathbf{u}^T \boldsymbol{\varphi}^{\alpha_\beta})\}\,. \tag{2.14}$$

Denoting then by $\phi_{\mathbf{H}}(\mathbf{v}) = E\{\exp(i\, \mathbf{v}^T \mathbf{H})\}$ the characteristic function of the random vector $\mathbf{H}$ leads finally to the equality

$$\psi_N(\mathbf{u}) = \exp(i\, \mathbf{u}^T \mathbf{m}_{\mathbf{X}}) \sum_{\alpha_1=1}^{m} \ldots \sum_{\alpha_N=1}^{m} \frac{\lambda_{\alpha_1} \ldots \lambda_{\alpha_N}}{\mu^N}\, \phi_{\mathbf{H}}(\sqrt{\frac{\mu}{N}}\, \mathbf{u}^T \boldsymbol{\varphi}^{\alpha_1}, \ldots, \sqrt{\frac{\mu}{N}}\, \mathbf{u}^T \boldsymbol{\varphi}^{\alpha_N})\,. \tag{2.15}$$

2.3.4 Taking the mathematical expectation of Eq. (2.4) it is directly found that $E\{\mathbf{X}^N_{\text{CPCA}}\} = \mathbf{m}_{\mathbf{X}}$. Moreover, one obtains

$$E\{(\mathbf{X}^N_{\text{CPCA}} - \mathbf{m}_{\mathbf{X}})\,(\mathbf{X}^N_{\text{CPCA}} - \mathbf{m}_{\mathbf{X}})^T\} = (\mu/N) \sum_{\alpha=1}^{N} \sum_{\beta=1}^{N} E\{H_\alpha\, H_\beta\, \boldsymbol{\varphi}^{J_\alpha}\, (\boldsymbol{\varphi}^{J_\beta})^T\} \tag{2.16}$$

$$= (\mu/N) \sum_{\alpha=1}^{N} E\{H_\alpha^2\}\, E\{\boldsymbol{\varphi}^{J_\alpha}\, (\boldsymbol{\varphi}^{J_\alpha})^T\} = \sum_{\alpha=1}^{m} \lambda_\alpha\, \boldsymbol{\varphi}^\alpha\, (\boldsymbol{\varphi}^\alpha)^T$$

using Eq. (2.12). However, the last expression in the above equation corresponds to the reconstruction of the covariance matrix $E\{(\mathbf{X}^m_{\text{PCA}} - \mathbf{m}_{\mathbf{X}})\,(\mathbf{X}^m_{\text{PCA}} - \mathbf{m}_{\mathbf{X}})^T\}$ from its eigenvalues and eigenvectors. This finding demonstrates that the compressed principal component approximation matches the first two moments of its standard PCA counterpart.

3. Expansion with a vector-valued symmetric polynomial chaos. To complete the representation of Eq. (2.4), it remains to characterize the random vector $\mathbf{H} = (H_1, \ldots, H_N)$. In the following sections, it is shown first (section 3.1) that, in approximating a given random vector $\mathbf{X}$ in the form of Eq. (2.4), it is sufficient to consider random vectors $\mathbf{H}$ with a probability density function $\mathbf{h} \mapsto p_{\mathbf{H}}(\mathbf{h})$ continuous on $\mathbb{R}^N$ that is a symmetric function with respect to $h_1, \ldots, h_N$, the components of $\mathbf{h}$. This means that for any permutation σ belonging to the set $\mathcal{S}_\sigma$ of the $N!$ permutations of integers $1, 2, \ldots, N$, we have $p_{\mathbf{H}}(\mathbf{h}_\sigma) = p_{\mathbf{H}}(\mathbf{h})$ in which $\mathbf{h}_\sigma = (h_{\sigma(1)}, \ldots, h_{\sigma(N)})$. This property could also be rewritten in terms of components as follows,

$$p_{\mathbf{H}}(h_{\sigma(1)}, \ldots, h_{\sigma(N)}) = p_{\mathbf{H}}(h_1, \ldots, h_N) \quad , \quad \forall\ \sigma \in \mathcal{S}_\sigma\,. \tag{3.1}$$

In the next section (section 3.2), a representation of random vectors $\mathbf{H}$ with symmetric probability density function is introduced in the form of

$$\mathbf{H} = \sum_{k=1}^{M} z_k\, \boldsymbol{\psi}^k(\boldsymbol{\Xi}) \tag{3.2}$$

in which $\boldsymbol{\psi}^k(\boldsymbol{\Xi})$ is a set of vector-valued symmetric polynomials depending on the canonical Gaussian random vector $\boldsymbol{\Xi} = (\Xi_1, \ldots, \Xi_N)$ and where the coefficients of the representation are gathered in vector $\mathbf{z} = (z_1, \ldots, z_M)$. The polynomials $\boldsymbol{\psi}^k(\boldsymbol{\Xi}) = (\psi^k_1(\boldsymbol{\Xi}), \ldots, \psi^k_N(\boldsymbol{\Xi}))$ are referred to as symmetric here to indicate the property

$$\boldsymbol{\psi}^k_\sigma(\boldsymbol{\Xi}) = (\psi^k_{\sigma(1)}(\boldsymbol{\Xi}), \ldots, \psi^k_{\sigma(N)}(\boldsymbol{\Xi})) = \boldsymbol{\psi}^k(\boldsymbol{\Xi}_\sigma) \quad , \quad \forall\ \sigma \in \mathcal{S}_\sigma\,. \tag{3.3}$$

Finally, in the last section (section 3.3), a construction of the symmetric polynomials $\boldsymbol{\psi}^k(\boldsymbol{\Xi})$ will be introduced.

3.1. Symmetry of the probability density function $p_{\mathbf{H}}(\mathbf{h})$. It is demonstrated in this section that the probability density function $p_{\mathbf{X}^N_{\text{CPCA}}}(\mathbf{x})$ obtained from Eq. (2.4) with random vectors $\mathbf{H}$ having probability density function $p^U_{\mathbf{H}}(\mathbf{h})$ or its symmetrized version

$$p_{\mathbf{H}}(\mathbf{h}) = \frac{1}{N!} \sum_{\sigma \in \mathcal{S}_\sigma} p^U_{\mathbf{H}}(\mathbf{h}_\sigma). \tag{3.4}$$

are identical. To prove this property, introduce first the $\mathbb{R}^n$-valued random variable $\mathbf{Y}$ as

$$\mathbf{Y} = \sqrt{\frac{N}{\mu}}\,(\mathbf{X}^N_{\text{CPCA}} - \mathbf{m}_{\mathbf{X}}) = \sum_{\alpha=1}^{N} H_\alpha\, \boldsymbol{\varphi}^{J_\alpha}, \tag{3.5}$$

Then, consider the contribution $p^p_{\mathbf{Y}}(\mathbf{y})$ to the probability density function $p_{\mathbf{Y}}(\mathbf{y})$ induced by one specific permutation $p^U_{\mathbf{H}}(\mathbf{h}_\sigma)$. It can be expressed as

$$p^p_{\mathbf{Y}}(\mathbf{y}) = E\{\delta_0(\widehat{\mathbf{y}}(\mathbf{H},\mathbf{J}) - \mathbf{y})\} \tag{3.6}$$

with $\mathbf{H}$ follows the probability measure $p^U_{\mathbf{H}}(\mathbf{h}_\sigma)\, d\mathbf{h}$, $\delta_0()$ denotes the Dirac measure at the origin, and

$$\widehat{\mathbf{y}}(\mathbf{H},\mathbf{J}) = \sum_{\alpha=1}^{N} H_\alpha\, \boldsymbol{\varphi}^{J_\alpha}. \tag{3.7}$$

Next, define the permutation $\widehat{\sigma}$ as the inverse of σ, i.e., such that $(\mathbf{H}_\sigma)_{\widehat{\sigma}} = \mathbf{H}$ and let $\widehat{\mathbf{H}} = \mathbf{H}_\sigma$ or $\mathbf{H} = \widehat{\mathbf{H}}_{\widehat{\sigma}}$. Then,

$$\widehat{\mathbf{y}}(\mathbf{H},\mathbf{J}) = \sum_{\alpha=1}^{N} \widehat{H}_{\widehat{\sigma}(\alpha)}\, \boldsymbol{\varphi}^{J_\alpha} = \sum_{\beta=1}^{N} \widehat{H}_\beta\, \boldsymbol{\varphi}^{J_{\sigma(\beta)}} = \sum_{\beta=1}^{N} \widehat{H}_\beta\, \boldsymbol{\varphi}^{\widehat{J}_\beta} = \widehat{\mathbf{y}}(\widehat{\mathbf{H}},\widehat{\mathbf{J}})\,, \tag{3.8}$$

where the second to last equality results from the change of index $J_{\sigma(\beta)} = \widehat{J}_\beta$. Substituting (3.8) into (3.6) yields

$$p^p_{\mathbf{Y}}(\mathbf{y}) = E\{\delta_0(\widehat{\mathbf{y}}(\widehat{\mathbf{H}},\widehat{\mathbf{J}}) - \mathbf{y})\}\,. \tag{3.9}$$

Then, comparing Eqs (3.6) and (3.9), it is concluded that the permutation σ has no effect on $p^p_{\mathbf{Y}}(\mathbf{y})$. Thus, the probability density function of $\mathbf{Y}$ determined from $p^U_{\mathbf{H}}(\mathbf{h}_\sigma)$ or its symmetrized version $p_{\mathbf{H}}(\mathbf{h})$ given in Eq. (3.4) are identical. From Eq. (3.5), one further concludes that the same property holds for the probability density function of $\mathbf{X}$.

3.2. Expansion of random vector H. On the basis of the symmetry of their probability density function, the random vectors $\mathbf{H}$ will be expressed in the form of Eq. (3.2) where the polynomials $\boldsymbol{\psi}^k(\boldsymbol{\Xi})$ are symmetric as defined by Eq. (3.3). In this section, it is shown that the corresponding probability density function of $\mathbf{H}$ is indeed symmetric. This function can be expressed as

$$p_{\mathbf{H}}(\mathbf{h}) = E\{\delta_0(\widetilde{\mathbf{h}}(\boldsymbol{\Xi}) - \mathbf{h})\}\,, \tag{3.10}$$

where

$$H = \widetilde{\mathbf{h}}(\boldsymbol{\Xi}) \quad , \quad \widetilde{\mathbf{h}}(\boldsymbol{\Xi}) = \sum_{k=1}^{M} z_k\, \boldsymbol{\psi}^k(\boldsymbol{\Xi})\,, \tag{3.11}$$

which, given Eq. (3.3), satisfies $\widetilde{\mathbf{h}}_\sigma(\boldsymbol{\Xi}) = \widetilde{\mathbf{h}}(\boldsymbol{\Xi}_\sigma)$. Using this property, Eq. (3.10) becomes

$$\begin{aligned} p_{\mathbf{H}}(\mathbf{h}_\sigma) &= E\{\delta_0(\widetilde{\mathbf{h}}(\boldsymbol{\Xi}) - \mathbf{h}_\sigma)\} = E\{\delta_0(\widetilde{\mathbf{h}}_{\widehat{\sigma}}(\boldsymbol{\Xi}) - \mathbf{h})\} \\ &= E\{\delta_0(\widetilde{\mathbf{h}}(\boldsymbol{\Xi}_{\widehat{\sigma}}) - \mathbf{h})\} = E\{\delta_0(\widetilde{\mathbf{h}}(\widehat{\boldsymbol{\Xi}}) - \mathbf{h})\} = p_{\mathbf{H}}(\mathbf{h})\,, \end{aligned} \tag{3.12}$$

where, as before, $\widehat{\sigma}$ denotes the inverse of the permutation σ. Moreover, the random vector is $\widehat{\boldsymbol{\Xi}} = \boldsymbol{\Xi}_{\widehat{\sigma}}$ and has the same distribution as $\boldsymbol{\Xi}$ as the Gaussian measure is symmetric.

Since Eq. (3.12) is valid for any permutation $\sigma \in S_\sigma$ the probability density function of $\mathbf{H}$ is symmetric.

3.3. Construction of the symmetric polynomials $\boldsymbol{\psi}^k(\boldsymbol{\xi})$. This section presents the construction of a novel set of symmetric vector-valued multidimensional polynomials $\boldsymbol{\psi}^k(\boldsymbol{\xi})$. To clarify this process, let any such symmetric polynomial homogeneous of order r in the variables $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$ be denoted as $\boldsymbol{\mathcal{Q}}^{(r)}(\boldsymbol{\xi}) = (\mathcal{Q}_1^{(r)}(\boldsymbol{\xi}), \ldots, \mathcal{Q}_N^{(r)}(\boldsymbol{\xi}))$. Expanding $\mathcal{Q}_\alpha^{(r)}(\boldsymbol{\xi})$ in terms of the component ξ_α yields

$$\mathcal{Q}_\alpha^{(r)}(\boldsymbol{\xi}) = \sum_{s=0}^{r} w_s^{(\alpha)}\, \xi_\alpha^s\; Q_\alpha^{(r-s)}(\boldsymbol{\xi}'^{(\alpha)}) \tag{3.13}$$

where $w_s^{(\alpha)}$ are parameters and $Q_\alpha^{(r-s)}(\boldsymbol{\xi}'^{(\alpha)})$ denotes a multidimensional polynomial of degree r–s in the N-1 component vector $\boldsymbol{\xi}'^{(\alpha)} = (\xi_1, \ldots, \xi_{\alpha-1}, \xi_{\alpha+1}, \ldots, \xi_N)$.

The next task is to establish the conditions on the parameters $w_s^{(\alpha)}$ and on the polynomials $Q_\alpha^{(r-s)}(\boldsymbol{\xi}'^{(\alpha)})$ so that $\boldsymbol{\mathcal{Q}}^{(r)}(\boldsymbol{\xi})$ is symmetric, i.e., $\boldsymbol{\mathcal{Q}}_\sigma^{(r)}(\boldsymbol{\xi}) = \boldsymbol{\mathcal{Q}}^{(r)}(\boldsymbol{\xi}_\sigma)$. To this end, consider first the ensemble of permutations σ' of $1, \ldots, \alpha-1, \alpha+1, \ldots, N$ that leave the component α untouched. The component α of $\boldsymbol{\mathcal{Q}}_{\sigma'}^{(r)}(\boldsymbol{\xi})$ is then

$$\{\boldsymbol{\mathcal{Q}}_{\sigma'}^{(r)}(\boldsymbol{\xi})\}_\alpha = \sum_{s=0}^{r} w_s^{(\alpha)}\, \xi_\alpha^s\; Q_\alpha^{(r-s)}(\boldsymbol{\xi}'^{(\alpha)}_{\sigma'}). \tag{3.14}$$

To satisfy the symmetry condition, this component should also equal $\sum_{s=0}^{r} w_s^{(\alpha)}\, \xi_\alpha^s\; Q_\alpha^{(r-s)}(\boldsymbol{\xi}'^{(\alpha)})$. For both expressions to be equal for any permutation σ', it is necessary and sufficient that $Q_\alpha^{(r-s)}(\boldsymbol{\xi}'^{(\alpha)})$ be a scalar symmetric polynomial in the variables $\boldsymbol{\xi}'^{(\alpha)}$.

Consider next permutations that affect the index α, which specifically is mapped from index β. Then, $\mathcal{Q}_\beta^{(r)}(\boldsymbol{\xi})$ must be mapped into $\mathcal{Q}_\alpha^{(r)}(\boldsymbol{\xi})$. Accordingly,

$$\sum_{s=0}^{r} w_s^{(\beta)}\, \xi_\alpha^s\; Q_\beta^{(r-s)}(\boldsymbol{\xi}'^{(\alpha)}) \text{ must equal } \sum_{s=0}^{r} w_s^{(\alpha)}\, \xi_\alpha^s\, Q_\alpha^{(r-s)}(\boldsymbol{\xi}'^{(\alpha)}). \tag{3.15}$$

The above condition is satisfied when neither the coefficients $w_s^{(\alpha)}$ nor the polynomials $Q_\alpha^{(s)}()$ depend on α for all s. Accordingly, these coefficients and polynomials will be denoted henceforth as w_s and $Q^{(s)}()$. With this finding, Eq. (3.13) becomes

$$\mathcal{Q}_\alpha^{(r)}(\boldsymbol{\xi}) = \sum_{s=0}^{r} w_s\, \xi_\alpha^s\; Q^{(r-s)}(\boldsymbol{\xi}'^{(\alpha)})\,, \tag{3.16}$$

The next step is the construction of the scalar symmetric polynomials $Q^{(s)}(\boldsymbol{\xi}'^{(\alpha)})$. We use the fact that any scalar symmetric polynomial can be represented using the power-sums (see [28]). On that basis, we propose to write the polynomial $Q^{(s)}(\boldsymbol{\xi}'^{(\alpha)})$ as

$$Q^{(s)}(\boldsymbol{\xi}'^{(\alpha)}) = \sum_t q_t^{(s)} \{ \prod_{u=1}^{N-1} (S_u(\boldsymbol{\xi}'^{(\alpha)}))^{p_t(u)} \}\,. \tag{3.17}$$

in which

$$S_u(\boldsymbol{\xi}'^{(\alpha)}) = \sum_{\ell=1}^{N-1} (\xi'^{(\alpha)}_\ell)^u \quad , \quad u = 1, \ldots, N-1 \tag{3.18}$$

are the power-sums and where $q_t^{(s)}$ are coefficients. Moreover, $p_t(u)$ are nonnegative integer powers such that

$$\sum_{u=1}^{N-1} u\, p_t(u) = s\,, \tag{3.19}$$

and the summation over t in Eq. (3.17) extends over all such possible combinations of powers. Recombining Eqs (3.16) and (3.17), yields

$$\mathcal{Q}_\alpha^{(r)}(\boldsymbol{\xi}) = \sum_{s=0}^{r} \sum_t y_t^{(r,s)}\, \xi_\alpha^s \prod_{u=1}^{N-1} (S_u(\boldsymbol{\xi}'^{(\alpha)}))^{p_t(u)}\,, \tag{3.20}$$

where $y_t^{(r,s)} = w_s q_t^{(r-s)}$ are parameters and the summation over t extends over all combinations of the powers $p_t(u)$ such that

$$\sum_{u=1}^{N-1} u\, p_t(u) + s = r \quad , \quad s \leq r\,. \tag{3.21}$$

The symmetric vector-valued multidimensional polynomials of component α equal to $\xi_\alpha^s \prod_{u=1}^{N-1} (S_u(\boldsymbol{\xi}'^{\alpha}))^{p_t(u)}$, corresponding to different values of r, s, and t are linearly independent and thus can be selected to form each of the desired polynomials $\boldsymbol{\psi}^k(\boldsymbol{\xi})$. Note however the condition $E\{H_\alpha\} = 0$ required by the model. To satisfy this condition, a constant c_k will be subtracted so that $E\{\psi_\alpha^k(\boldsymbol{\Xi})\} = 0$. That is,

$$\psi_\alpha^k(\boldsymbol{\xi}) = \xi_\alpha^s \prod_{u=1}^{N-1} (S_u(\boldsymbol{\xi}'^{\alpha}))^{p_t(u)} - c_k\,. \tag{3.22}$$

Note that the constant c_k corresponds in fact to the 0^{th}-order polynomial ignored in the previous developments.

The number M of symmetric vector-valued polynomials $\{\boldsymbol{\psi}^k(\boldsymbol{\Xi}), k = 1, \ldots, M\}$ in Eq. (3.22) is shown in Table 1 as function of N and n_d, the maximum order considered. Observe in particular that this number remains constant for $N \geqslant n_d + 1$ and is further independent of both n and m, the number of components in the vector $\mathbf{X}$ and the number of PCA eigenvectors retained.

As examples, shown below are the α components, $\alpha = 1, \ldots, N$, of the symmetric vector-valued multidimensional polynomials of degree $r = 1$ and 2 (for $N > 2$)

$$\psi_\alpha^1(\boldsymbol{\xi}) = \xi_\alpha \quad , \quad \psi_\alpha^2(\boldsymbol{\xi}) = \sum_{\beta=1, \beta\neq\alpha}^{N} \xi_\beta\,, \tag{3.23}$$

$$\psi_\alpha^3(\boldsymbol{\xi}) = \xi_\alpha^2 - 1 \quad , \quad \psi_\alpha^4(\boldsymbol{\xi}) = \xi_\alpha \sum_{\beta=1, \beta\neq\alpha}^{N} \xi_\beta\,, \tag{3.24}$$

$$\psi_\alpha^5(\boldsymbol{\xi}) = \left(\sum_{\beta=1, \beta\neq\alpha}^{N} \xi_\beta \right)^2 - (N-1) \quad , \quad \psi_\alpha^6(\boldsymbol{\xi}) = \sum_{\beta=1, \beta\neq\alpha}^{N} (\xi_\beta^2 - 1)\,. \tag{3.25}$$

TABLE 1

Table 1. Number M of vector symmetric polynomials vs. the number of eigenvalues (N) and the polynomial maximum degree (n_d)

N	n_d						
	1	2	3	4	5	6	7
2	2	5	9	14	20	27	35
3	2	6	12	21	33	49	69
4	2	6	13	24	40	63	94
5	2	6	13	25	43	70	108
6	2	6	13	25	44	73	115
7	2	6	13	25	44	74	118
8	2	6	13	25	44	74	119
9	2	6	13	25	44	74	119

Two matlab functions are presented as supplementary material that construct the set of vector-valued polynomial given input values of N, n_d, and samples $\boldsymbol{\xi}$. These functions should be called consecutively as
[Rpoly,MatRpoly]=sub_scal_symm_poly(N-1,Nd)
[Npoly,MatRPsi]=sub_symm_poly_eval(N,Nd,Rpoly,MatRpoly,MatRxi)
see instructions inside the functions.

4. Identification of the parameters. The focus of the CPCA approach is on modeling a random vector $\mathbf{X}$ with n_e available independent realizations $\mathbf{x}^{(\ell)}$, $\ell = 1, \ldots, n_e$, referred to as "experiments" in the sequel, in the form of Eqs (2.4) and (3.2). This modeling involves the determination of estimates (denoted by an overlined hat) of the mean $\mathbf{m}_{\mathbf{X}}$, the eigenvalues λ_α, the corresponding eigenvector $\boldsymbol{\varphi}^\alpha$, and the parameters z_k of Eq. (3.2).

The first three quantities will be estimated using the sample mean and covariance matrix, i.e.,

$$\widehat{\mathbf{m}}_{\mathbf{X}} = \frac{1}{n_e} \sum_{\ell=1}^{n_e} \mathbf{x}^{(\ell)} \tag{4.1}$$

$$[\widehat{C}_{\mathbf{X}}] = \frac{1}{n_e - 1} \sum_{\ell=1}^{n_e} (\mathbf{x}^{(\ell)} - \widehat{\mathbf{m}}_{\mathbf{X}}) \, (\mathbf{x}^{(\ell)} - \widehat{\mathbf{m}}_{\mathbf{X}})^T \tag{4.2}$$

Then, the solution of the eigenvalue problem

$$[\widehat{C}_{\mathbf{X}}] \, \widehat{\boldsymbol{\varphi}}^\alpha = \widehat{\lambda}_\alpha \, \widehat{\boldsymbol{\varphi}}^\alpha \tag{4.3}$$

yields the required estimates of the PCA eigenvectors, $\widehat{\boldsymbol{\varphi}}^\alpha$, and eigenvalues $\widehat{\lambda}_\alpha$.

To complete the identification procedure, it then remains to estimate the coefficients z_k of Eq. (3.2). Several standard strategies are available for this task such as the maximum likelihood approach. Other methods are based on the minimization of a distance between the distributions induced by the model and by the measurements. In this paper, two identification approaches are proposed.

4.1. Direct minimization of the overlap error. The first identification method is based on a direct minimization of the overlap error defined in Eqs (2.7) and (2.8)

which is written as

$$\mathbf{z}^{\rm opt} = \arg\min_{\mathbf{z}} \epsilon_{ovl}(\mathbf{z})\,, \tag{4.4}$$

where $\mathbf{z}$ is required to satisfy the constraints

$$E\{H_\alpha H_\beta\} = \delta_{\alpha\beta}\,. \tag{4.5}$$

Given the symmetry of the vector-valued multidimensional polynomials $\boldsymbol{\psi}^k(\boldsymbol{\Xi})$, the resulting conditions are identical for all components α (unit variance) or pair (α,β) of distinct components (uncorrelatedness). Accordingly, these conditions can be averaged separately over all values $\alpha=\beta$ and $\alpha\neq\beta$ yielding only two constraints. In terms of n_e realized values, these constraints can be written as

$$\frac{1}{N\,n_e}\sum_{\ell=1}^{n_e}\mathbf{h}^{(\ell)T}\,\mathbf{h}^{(\ell)} = 1\,, \tag{4.6}$$

and

$$\frac{1}{n_e(N^2-N)}\sum_{\ell=1}^{n_e}\sum_{\alpha=1}^{N}\sum_{\substack{\beta=1\\ \beta\neq\alpha}}^{N} h_\alpha^{(\ell)}\,h_\beta^{(\ell)} = 0\,. \tag{4.7}$$

With Eq. (3.2), these constraints can be rewritten as

$$\mathbf{z}^T\,[A]\,\mathbf{z} = 1 \;\text{ and }\; \mathbf{z}^T\,[B]\,\mathbf{z} = 0\,, \tag{4.8}$$

where the vector $\mathbf{z}=(z_1,\ldots,z_M)$ and the matrices $[A]$ and $[B]$ have components

$$[A]_{kk'} = \frac{1}{N\,n_e}\sum_{\ell=1}^{n_e}\boldsymbol{\psi}^k(\widehat{\boldsymbol{\xi}}^{(\ell)})^T\,\boldsymbol{\psi}^{k'}(\widehat{\boldsymbol{\xi}}^{(\ell)})\,, \tag{4.9}$$

$$[B]_{kk'} = \frac{1}{n_e(N^2-N)}\sum_{\ell=1}^{n_e}\sum_{\alpha=1}^{N}\sum_{\substack{\beta=1\\ \beta\neq\alpha}}^{N}\psi_\alpha^k(\widehat{\boldsymbol{\xi}}^{(\ell)})\,\psi_\beta^{k'}(\widehat{\boldsymbol{\xi}}^{(\ell)})\,. \tag{4.10}$$

One challenge associated with this effort is its computational cost. Indeed, at each iteration it is necessary to estimate the distribution of each element of the random vector $\mathbf{X}_{\rm CPCA}^N$. Since n, the number of these components, may be very large, the computational cost may be excessive. Thus, an alternate approach is also proposed in the next section.

4.2. Least square identification. In this section, a least squares identification of the coefficients z_k is proposed as

$$\mathbf{z}^{\rm opt} = \arg\min_{\mathbf{z}} \epsilon_{\rm LS}(\mathbf{z})\,, \tag{4.11}$$

where $\epsilon_{\rm LS}(\mathbf{z})$ is either of $\epsilon_{\mathbf{X}}(\mathbf{z})$ or $\epsilon_{\mathbf{H}}(\mathbf{z})$ defined below and satisfying the constraints of Eqs (4.8)-(4.10). The first error is

$$\epsilon_{\mathbf{X}}(\mathbf{z}) = E\left\{\left\|\mathbf{X}-\mathbf{m}_{\mathbf{X}} - \sqrt{\frac{\mu}{N}}\sum_{\alpha=1}^{N}\sum_{k=1}^{M} z_k\,\boldsymbol{\psi}_\alpha^k(\boldsymbol{\Xi})\,\boldsymbol{\varphi}^{J_\alpha}\right\|^2\right\}\,, \tag{4.12}$$

which is estimated by

$$\widehat{\epsilon}_{\mathbf{X}}(\mathbf{z}) = \frac{1}{n_e} \sum_{\ell=1}^{n_e} \left\| \mathbf{x}^{(\ell)} - \widehat{\mathbf{m}}_{\mathbf{X}} - \sqrt{\frac{\mu}{N}} \sum_{\alpha=1}^{N} \sum_{k=1}^{M} z_k \, \boldsymbol{\psi}_{\alpha}^{k}(\boldsymbol{\xi}^{(\ell)}) \, \widehat{\boldsymbol{\varphi}}^{j_{\alpha}^{(\ell)}} \right\|^2 , \tag{4.13}$$

where $\boldsymbol{\xi}^{(\ell)}$ and $\mathbf{j}^{(\ell)}$ are the ℓ^{th} realizations of the random vectors $\boldsymbol{\Xi}$ and $\mathbf{J}$ associated with $\mathbf{x}^{(\ell)}$. The second error is

$$\epsilon_{\mathbf{H}}(\mathbf{z}) = E\left\{ \left\| \mathbf{H} - \sum_{k=1}^{M} z_k \, \boldsymbol{\psi}^k(\boldsymbol{\Xi}) \right\|^2 \right\} , \tag{4.14}$$

which is estimated by

$$\widehat{\epsilon}_{\mathbf{H}}(\mathbf{z}) = \frac{1}{n_e} \sum_{\ell=1}^{n_e} \left\| \mathbf{h}^{(\ell)} - \sum_{k=1}^{M} z_k \, \boldsymbol{\psi}^k(\boldsymbol{\xi}^{(\ell)}) \right\|^2 \tag{4.15}$$

where $\mathbf{h}^{(\ell)}$ denotes the ℓ^{th} realization of the random vector $\mathbf{H}$ associated with $\mathbf{x}^{(\ell)}$. The above strategies appear computationally efficient but they require the experimental values $\boldsymbol{\xi}^{(\ell)}$ and $\mathbf{j}^{(\ell)}$, which are not available. This issue is resolved below considering the error $\widehat{\epsilon}_{\mathbf{H}}(\mathbf{z})$ of Eq. (4.15). Similar steps can be followed when considering the error $\widehat{\epsilon}_{\mathbf{X}}(\mathbf{z})$ of Eq. (4.13).

STEP 1: In the absence of experimental values for $\mathbf{J}$, a set of independent realizations $\mathbf{j}^{(\ell)}$ are generated according to the distribution of Eqs (2.5) and (2.6) with λ_α replaced by $\widehat{\lambda}_\alpha$.

STEP 2: Estimates of the corresponding experimental values of $\widehat{\mathbf{h}}^{(\ell)} = (\widehat{h}_1^{(\ell)}, \ldots, \widehat{h}_N^{(\ell)})$ are obtained for each $\ell = 1, \ldots, n_e$ by minimizing the error

$$\epsilon_{\text{CPCA}}^{(\ell)} = \left\| \mathbf{x}^{(\ell)} - \widehat{\mathbf{m}}_{\mathbf{X}} - \sqrt{\frac{\mu}{N}} \sum_{\alpha=1}^{N} \widehat{h}_{\alpha}^{(\ell)} \, \widehat{\boldsymbol{\varphi}}^{j_{\alpha}^{(\ell)}} \right\|^2 . \tag{4.16}$$

This optimization effort leads to the linear system of equations

$$[\widehat{\boldsymbol{\phi}}^{(\ell)}]^T [\widehat{\boldsymbol{\phi}}^{(\ell)}] \, \widehat{\mathbf{h}}^{(\ell)} = \sqrt{\frac{N}{\mu}} \, [\widehat{\boldsymbol{\phi}}^{(\ell)}]^T \, (\mathbf{x}^{(\ell)} - \widehat{\mathbf{m}}_{\mathbf{X}}) , \tag{4.17}$$

where

$$[\widehat{\boldsymbol{\phi}}^{(\ell)}] = [\widehat{\boldsymbol{\varphi}}^{j_1^{(\ell)}} \ldots \widehat{\boldsymbol{\varphi}}^{j_N^{(\ell)}}] , \tag{4.18}$$

and from which n_e realizations of the vector $\mathbf{H}$ are obtained. Note that the empirical mean constructed from the identified values of $\widehat{\mathbf{h}}^{(\ell)}$ is enforced to be zero by removing it from the identified values.

STEP 3: The next step of the process is the construction of experimental values $\widehat{\boldsymbol{\xi}}^{(\ell)}$, i.e., realizations of the Gaussian random vector $\boldsymbol{\Xi}$ with covariance matrix equal to the identity matrix, but which have to be statistically dependent of $\widehat{\mathbf{h}}^{(\ell)}$ so that the least squares minimization problem of Eq. (4.15) is well posed. To ensure this dependence, the values $\widehat{\boldsymbol{\xi}}^{(\ell)}$ will be determined from $\widehat{\mathbf{h}}^{(\ell)}$ in an iterative process seeking

to approximate at best the independence and Gaussian properties of the random variables $\{\Xi_\alpha\}_\alpha$.

Specifically, denote by $[G^k]$ a $n \times n_e$ matrix at iteration k. It is intended that $[G^1] = [\widehat{\mathbf{h}}^{(1)} \dots \widehat{\mathbf{h}}^{(n_e)}]$ and at convergence of the sequence $\{[G^k]\}_k$, the columns of $[G^k]$ are n_e independent realizations $[\widehat{\boldsymbol{\xi}}^{(1)} \dots \widehat{\boldsymbol{\xi}}^{(n_e)}]$ of the Gaussian vector $\boldsymbol{\Xi}$. A transformation of matrix $[G^k]$ is first achieved to create uncorrelatedness. That is,

$$[G^{k+1}] = [L^k]^{-1} [G^k] \tag{4.19}$$

where the matrix $[L^k]$ is the decomposition (e.g., Cholesky) of the correlation matrix $[C_{[G^k]}]$, i.e.,

$$[L^k] [L^k]^T = [C_{[G^k]}] = \frac{1}{n_e - 1} [G^k] [G^k]^T. \tag{4.20}$$

The next sub-iteration maps each component of $[G^{k+1}]_{\alpha\ell}$ to the corresponding component of $[G^{k+2}]_{\alpha\ell}$ to achieve Gaussianity of each variable Ξ_α. Specifically,

$$[G^{k+2}]_{\alpha\ell} = F_G^{-1}(F_\alpha([G^{k+1}]_{\alpha\ell})) \tag{4.21}$$

where $F_G(.)$ and $F_\alpha(.)$ denote the cumulative distribution functions of the normal distribution and the one corresponding to the samples $[G^{k+1}]_{\alpha\ell}$, $\ell = 1, \dots, n_e$, for each α. Note that the exponent -1 indicates here the inverse function.

STEP 4: Having constructed samples $\widehat{\boldsymbol{\xi}}^{(\ell)}$, the entire set of polynomials $\boldsymbol{\psi}^k$ can be constructed according to Eq. (3.22). The constants c_k are evaluated so that the empirical mean of $\psi_\alpha^k(\boldsymbol{\Xi})$ estimated using $\widehat{\boldsymbol{\xi}}^{(\ell)}$ as realizations of $\boldsymbol{\Xi}$ is zero for every polynomial.

STEP 5: The imposition of the constraints of Eqs (4.8)-(4.10) is carried out using the Lagrange multiplier approach leading to the system of equations

$$\left(\sum_{\ell=1}^{n_e} [\boldsymbol{\Psi}^{(\ell)}]^T [\boldsymbol{\Psi}^{(\ell)}] + \Lambda_1 [A] + \Lambda_2 [B] \right) \mathbf{z} = \sum_{\ell=1}^{n_e} [\boldsymbol{\Psi}^{(\ell)}]^T \widehat{\mathbf{h}}^{(\ell)} \tag{4.22}$$

where

$$[\boldsymbol{\Psi}^{(\ell)}] = [\boldsymbol{\psi}^1(\widehat{\boldsymbol{\xi}}^{(\ell)}) \dots \boldsymbol{\psi}^M(\widehat{\boldsymbol{\xi}}^{(\ell)})] \tag{4.23}$$

and Λ_1 and Λ_2 are the Lagrange multipliers. The determination of the parameters z_k is most easily carried out by viewing the solution of Eq. (4.22) as $\mathbf{z} = \mathbf{z}(\Lambda_1, \Lambda_2)$. Then, Eq. (4.8) appear as 2 nonlinear algebraic equations for Λ_1 and Λ_2 the solution of which first yields these multipliers and then the corresponding vector $\mathbf{z}$.

STEP 6: The last step of the identification is the estimation of the error associated with the CPCA modeling. It is proposed here to quantify it using the overlap measures of Eqs (2.7) and (2.8). The nonparametric estimation of $p_{X_i}(x)$ based on the use of Eqs (2.4) and (3.2) is achieved with new realizations $\boldsymbol{\xi}^{(\ell)}$ and $\mathbf{j}^{(\ell)}$. The number of simulations, denoted here as n_s, should be selected large enough to obtain a reliable estimate of the probability density functions $p_{X_i}(x)$ and regardless of the value of n_e.

STEP 7: Since the least squares approach does not optimize Eqs (2.7) and (2.8) but rather Eq. (4.15), it is beneficial to repeat the above process for several values of N and n_d and then select the optimum CPCA approximation as the one yielding the lowest overlap error.

5. Numerical illustration. The construction of a CPCA representation was achieved for a non-Gaussian random vector constructed from the sampling in space of a 3-dimensional zero mean homogeneous real-valued Gaussian process $Z(\mathbf{s})$, $\mathbf{s} = (s_1, s_2, s_3)$ indexed by $\mathbb{R}^3$, with autocorrelation function

$$R_{ZZ}(\bar{\mathbf{s}}) = E[Z(\mathbf{s})Z(\mathbf{s}+\bar{\mathbf{s}})] = \prod_{\alpha=1}^{3} \frac{sin^2(p_\alpha)}{p_\alpha^2} \tag{5.1}$$

where

$$p_\alpha = (\pi \bar{s}_\alpha)/(2L_\alpha),\ \alpha = 1, 2, 3 \tag{5.2}$$

in which L_α denotes the correlation length in direction α.

The process $Z(\mathbf{s})$ was sampled at 9 locations in each direction with coordinates $\bar{s}_\alpha = 0, 0.125, 0.25, \ldots, 1$ and stacked to form the $n = 9^3 = 729$ component zero mean random vector $\mathbf{X}^G$ which is fully defined by its covariance matrix

$$\mathbf{K}_{\mathbf{X}^G\mathbf{X}^G} = E[\mathbf{X}^G(\mathbf{X}^G)^T] \tag{5.3}$$

The element $\alpha\beta$ of this matrix is

$$[\mathbf{K}_{\mathbf{X}^G\mathbf{X}^G}]_{\alpha\beta} = R_{ZZ}(\bar{\mathbf{s}}^{(\beta)} - \bar{\mathbf{s}}^{(\alpha)}) \tag{5.4}$$

where $\bar{\mathbf{s}}^{(\alpha)}$ is the 3-dimensional vector of the coordinates of the component α of $\mathbf{X}^G$.

The generation of samples of the Gaussian random vector $\mathbf{X}^G$ is easily achieved by first proceeding with a Cholesky decomposition of the positive-definite matrix $\mathbf{K}_{\mathbf{X}^G\mathbf{X}^G}$ as $\mathbf{K}_{\mathbf{X}^G\mathbf{X}^G} = \mathbf{L}\mathbf{L}^T$. Then, a component α of the random vector $\mathbf{X}^G$ can be expressed as

$$X_\alpha^G = \sum_{\beta=1}^{729} L_{\alpha\beta}\, \bar{\Xi}_\beta \tag{5.5}$$

where $\bar{\Xi}_\beta$, $\beta = 1, \ldots, 729$ are independent zero mean unit variance Gaussian random variables.

The non-Gaussian process of interest here will be expressed similarly to Eq. (5.5) as

$$X_\alpha = \sum_{\beta=1}^{729} L_{\alpha\beta}\, \bar{\Xi}_\beta^3. \tag{5.6}$$

For the results presented below, the correlation lengths were selected as $L_\alpha = 1$ and $n_e = 20\,000$ samples of the vector $\mathbf{X}$ were simulated and utilized. This large number of samples was selected to insure a converged behavior of the CPCA modeling process permitting the clear assessment of the methodology. The convergence was confirmed by increasing further n_e up to 160 000. Then, shown in Figs 1 and 2 are the PCA eigenvalues λ_i and the corresponding PCA error ϵ_{PCA} plotted vs. the number of PCA eigenvalues m for the non-Gaussian process of Eq. (5.6). To exemplify the CPCA construction, the approximation corresponding to $m = 15$ is first selected. It corresponds to a low error ϵ_{PCA}=0.037. The CPCA modeling was performed using both the direct minimization of the error of Eqs (2.7) and (2.8) and the least squares approach of Eq. (4.22) for various values of N and the polynomial order n_d for both

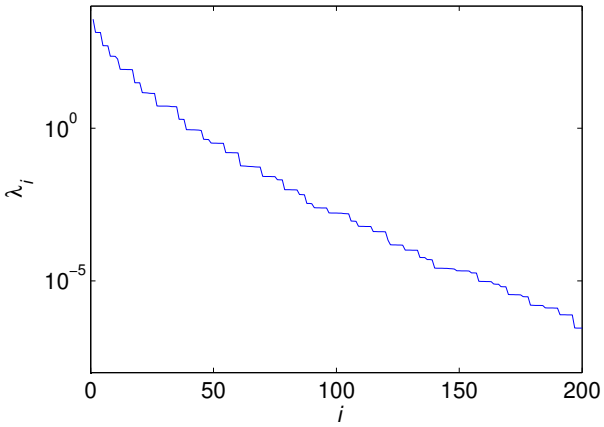

FIG. 1. *PCA Eigenvalues.*

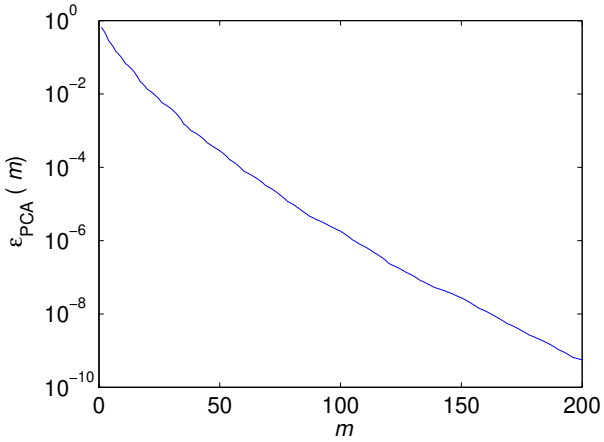

FIG. 2. *Error* $\epsilon_{\mathrm{PCA}}(m)$ *vs. number of eigenvectors* m.

methods. Note that the error $\epsilon_{ovl}(N, n_d)$ was obtained with the number of simulation n_s selected as 20 000 and the Gaussian kernel approach [6, 53] was used to estimate all probability density functions.

The direct minimization of the error was achieved first and shown in Fig. 3 are the corresponding minimum values obtained for various values of N and $n_d \geq 2$ starting with the common initial condition $z_1 = 1$ and all other coefficients z_k set to zero. The case $n_d = 1$, not shown here for clarity, which corresponds to the random components H_α being independent standard Gaussian random variables achieves a minimum overlap error of 0.21. It is seen in Fig. 3 that the error decreases as a function of N and n_d but achieves near convergence for $n_d = 3$ and $N = 3$. The increase in the overlap error occurring between $n_d = 3$ and $N = 3$ and $n_d = 4$ and $N = 3$ is believed to be associated with a local minimum which the minimization process was found to exhibit. Moreover, note that the solutions corresponding to $n_d = 2$ are definitely below the 0.21 value obtained for $n_d = 1$ demonstrating the role of the quadratic polynomials even though the distribution of the components of the

random vector $\mathbf{X}$ is odd, i.e., $p_{\mathbf{X}}(\mathbf{x}) = p_{\mathbf{X}}(-\mathbf{x})$ given its construction of Eq. (5.6). Also shown in Fig. 3 is the least squares solution, which was found to lead to a minimal overlap error equal to 0.164 and also occurring for $n_d = 3$ and $N = 3$. As a perspective for the results of Fig. 3, note that a Gaussian modeling of the random vector $\mathbf{X}$ would lead to an overlap error of 0.303.

The smoothed histograms of the 729 values of $\epsilon_{ovl}^{(i)}$, see Fig. 4, show that the accuracy of the CPCA approximations varies from one component to another. In particular, note that the peak of this histogram occurs for $\epsilon_{ovl}^{(i)} = 0.06-0.07$ depending on the method used even though the norm $\epsilon_{ovl}^{(i)} = 0.140 - 0.164$ owing to the long tail shown in Fig. 4.

Also shown on this figure is the error $\epsilon_{ovl}^{(i)}$, which would be obtained by modeling each component X_α of the random vector $\mathbf{X}$ as a Gaussian random variable. Even though the CPCA approximations involve only 6 (for $n_d = 2$ and $N = 3$) or 12 (for $n_d = 3$ and $N = 3$) coefficients, see Table 1, they yield a dramatic reduction of the overlap error in comparison to the Gaussian approximation.

A more detailed perspective on the modeling of each component can be obtained by comparing the marginal probability density functions of typical components X_α of the random vector $\mathbf{X}$. This comparison is shown in Figs 5-10 for some typical components for the least squares solution corresponding to $n_d = 3$ and $N = 3$. The matching obtained with the two approximations, $n_d = 2$ or 3 and $N = 3$, obtained with direct minimization is very similar for similar values of the overlap error $\epsilon_{ovl}^{(i)}$. In particular, no obvious asymmetry was noted for either approximation. The components chosen for Figs 5-10 are α=157, 329, and 129 to which correspond the errors $\epsilon_{ovl}^{(i)}$=0.0251, 0.0643, and 0.1654, respectively. For the first two cases, the matching between the CPCA distribution and its simulated ("Experiment") counterpart is excellent, both near the peak and in the tail. For the last case, the peak is not well captured but the agreement in the tail is still excellent. These results are a dramatic improvement over the Gaussian approximations also presented on these figures which match neither near the peak nor in the tail. Also presented in Figs. 5, 7, and 9 are the probability density functions obtained for the corresponding PCA approximations $X_{\text{PCA},\alpha}^{15}$ (curve "Truncation") which match closely their simulation counterparts demonstrating that the choice of m=15 eigenvectors is an appropriate selection. To complement the above analysis, shown in Figs 11 and 12 are the identical marginal distributions of the components H_α of the random vector $\mathbf{H}$ corresponding to the optimum selections $N = 3$, $n_d = 2$ and $n_d = 3$ as well as the least squares solution $N = 3$ and $n_d = 3$. Note in this figure the clear asymmetry corresponding to the $n_d = 2$ case and induced by the quadratic polynomials.

Finally, it was of interest to assess the CPCA approach for various values of m. This effort is summarized in Fig. 13 showing the minimum (over N for n_d=3) value of ϵ_{ovl} obtained with both the least squares and direct overlap minimization approaches. Consistently with Fig. 3, it is seen that the least squares approach leads to only small increases in error as compared to the direct overlap minimization. Also shown on this figure are the corresponding errors obtained for the PCA approximation as well as for a Gaussian modeling of each component independently. For small values of m, the error is primarily due to the PCA truncation but the resulting components are approximately Gaussian so that both the Gaussian approximation and the CPCA lead to similar errors. As m increases, the truncation error decreases and the components of the resulting random vector exhibit a more non-Gaussian (sharper peak) behavior. These trends are conflicting for the Gaussian approximation, which

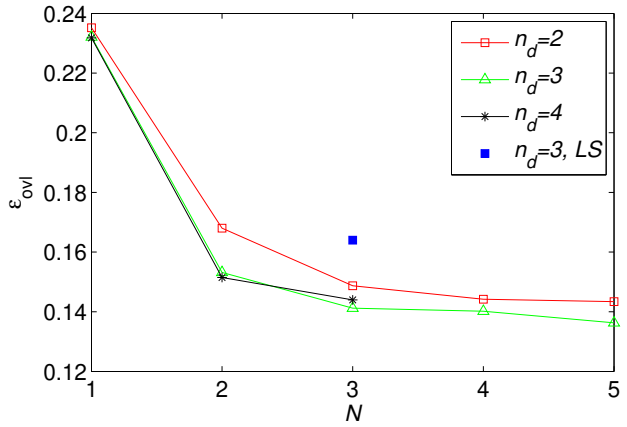

FIG. 3. *Minimum error $\epsilon_{ovl}(N, n_d)$ vs. polynomial order n_d and number of terms N, m=15, various optima with direct minimization and least squares approach ("LS")*

first decreases, reflecting the decrease in the truncation error, but then increases owing to the increased non-Gaussian character of the components X_α. This is not the case for the CPCA approach, that leads to an error ϵ_{ovl} which decreases monotonically stabilizing for m=9.

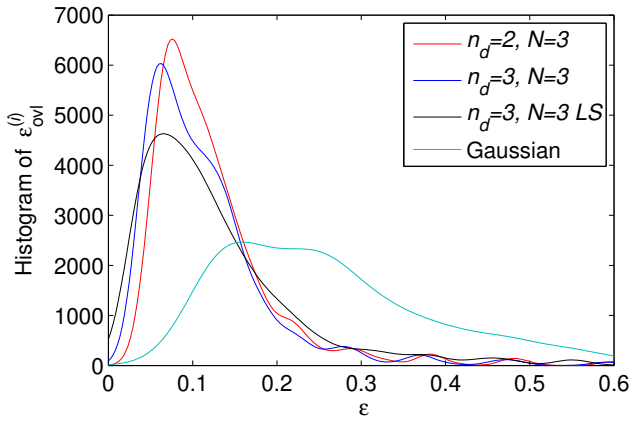

FIG. 4. *Smoothed histograms of the errors $\epsilon^{(i)}_{ovl}$, m=15, various optima with direct minimization and least squares approach ("LS") as well as for a Gaussian approximation.*

6. Discussion and conclusions. The present investigation has introduced a new approximate representation of non-Gaussian random vectors $\mathbf{X}$, referred to as Compressed Principal Component Analysis (CPCA), as a linear combination of a random sample of N eigenvectors of the PCA, see Eq. (2.4). In this representation, the random indices J_α of the eigenvectors are independent and have a distribution specified in terms of the PCA eigenvalues, see Eq. (2.6). The random coefficients H_α of the linear combination are by definition zero mean, uncorrelated, have unit variance, and are independent of the random indices J_α. Even without further characterization of the distribution of these coefficients, it was shown that the first and second order

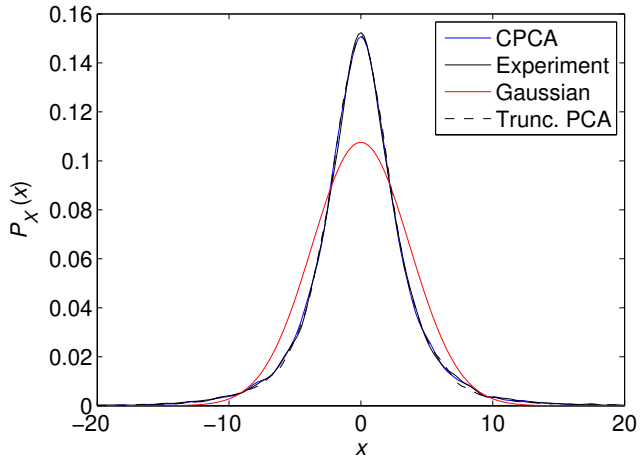

FIG. 5. *Probability density functions $p_X(x)$, of variable X_{157}, linear scale. Original data ("Experiment"), PCA truncated to m=15 terms ("Trunc. PCA"), its Gaussian approximation, and CPCA optimum least squares solution with N=3 and n_d=3, $\epsilon_{ovl}^{(157)}$=0.0251.*

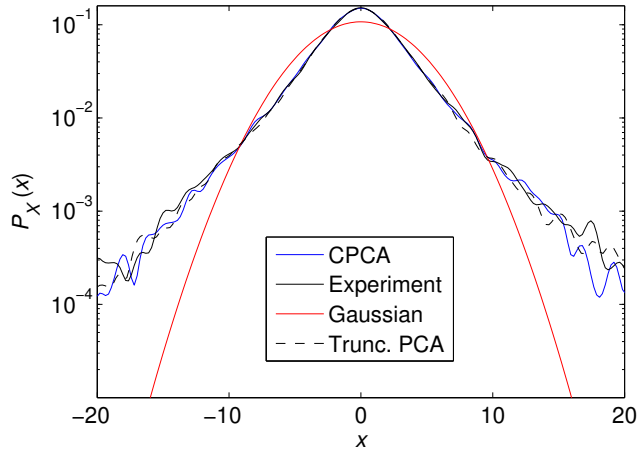

FIG. 6. *Probability density functions $p_X(x)$, of variable X_{157}, logarithmic scale. Original data ("Experiment"), PCA truncated to m=15 terms ("Trunc. PCA"), its Gaussian approximation, and CPCA optimum least squares solution with N=3 and n_d=3, $\epsilon_{ovl}^{(157)}$=0.0251.*

moments of the CPCA approximation are equal to those of the corresponding PCA representation.

A key property demonstrated next is that the probability distribution of the random coefficients H_α can be taken, without loss of generality, to be a symmetric function. This rather unique property allows to represent the vector of these coefficients in terms of a small number of deterministic parameter, z_k, in its expansion on a novel set $\{\boldsymbol{\psi}^k(\boldsymbol{\Xi})\}_k$ of symmetric vector-valued multidimensional polynomial of the canonical Gaussian random vector $\boldsymbol{\Xi}$, see Eqs (3.2) and (3.22)-(3.25).

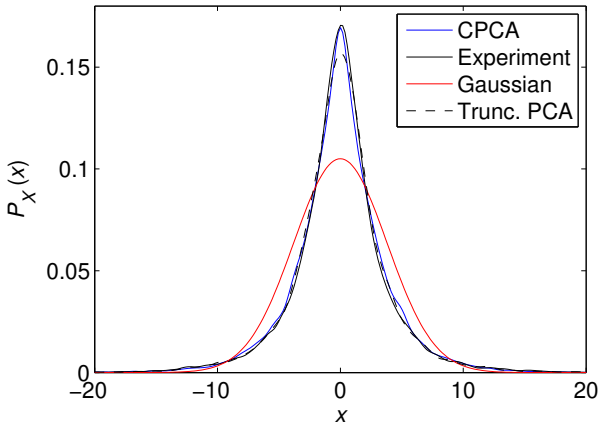

FIG. 7. *Probability density functions $p_X(x)$, of variable X_{329}, linear scale. Original data ("Experiment"), PCA truncated to m=15 terms ("Trunc. PCA"), its Gaussian approximation, and CPCA optimum least squares solution with N=3 and n_d=3, $\epsilon_{ovl}^{(329)}$=0.0643.*

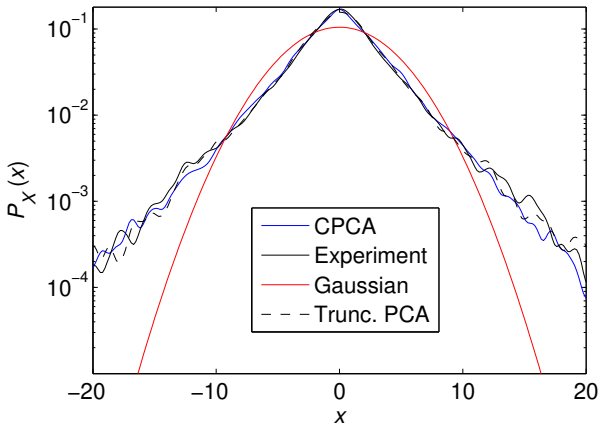

FIG. 8. *Probability density functions $p_X(x)$, of variable X_{329}, logarithmic scale. Original data ("Experiment"), PCA truncated to m=15 terms ("Trunc. PCA"), its Gaussian approximation, and CPCA optimum least squares solution with N=3 and n_d=3, $\epsilon_{ovl}^{(329)}$=0.0643.*

The identification of the parameters z_k from realizations of the random vector $\mathbf{X}$ was addressed next. A first optimization problem was proposed, i.e., the minimization of the overlap error of Eqs (2.7) and (2.8), which measures the difference between the distributions of the components X_α as determined from the available data and as modeled from the CPCA representation. Since finding the solution of this optimization problem could be computationally expensive, a much cheaper, sub-optimum alternative was also proposed, i.e., the minimization of either of the two least squares errors of Eqs (4.14)-(4.16). The validation example has confirmed that this least squares identification approach is indeed sub-optimum with respect to the overlap error but nevertheless provides a good approximation of the optimum solution obtained by a direct optimization of the overlap error. Moreover, the least squares

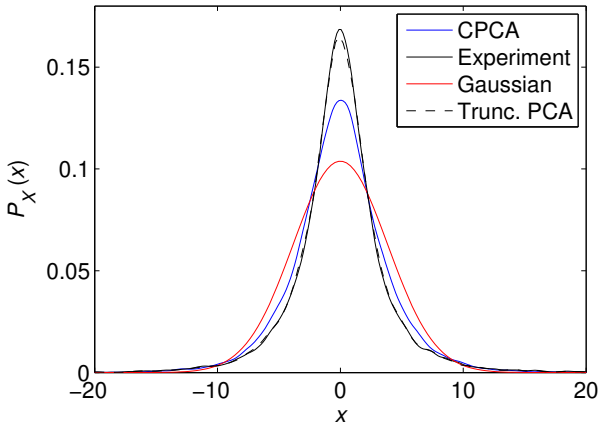

FIG. 9. *Probability density functions $p_X(x)$, of variable X_{129}, linear scale. Original data ("Experiment"), PCA truncated to m=15 terms ("Trunc. PCA"), its Gaussian approximation, and CPCA optimum least squares solution with N=3 and n_d=3, $\epsilon_{ovl}^{(129)}$=0.1654.*

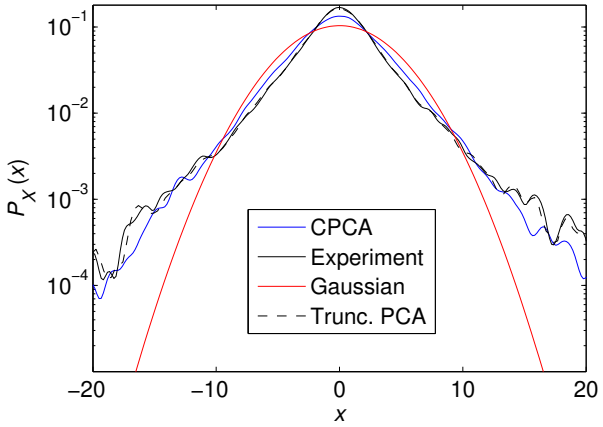

FIG. 10. *Probability density functions $p_X(x)$, of variable X_{129}, logarithmic scale. Original data ("Experiment"), PCA truncated to m=15 terms ("Trunc. PCA"), its Gaussian approximation, and CPCA optimum least squares solution with N=3 and n_d=3, $\epsilon_{ovl}^{(129)}$=0.1654.*

solution is obtained at a much reduced computational cost, which does not increase notably with increasing size of the random vector $\mathbf{X}$ and is thus applicable to large computational problems.

As discussed in Section 2.3, the CPCA representation is in general an approximation of its PCA counterpart. This is certainly expected when the number $2N$ of random variables it involves is less than the PCA order m. However, this is also the case when $2N > m$ as it is not expected that the characteristic function of the modeled vector, see Eq. (2.15), can match exactly any given characteristic function. Nevertheless, the strengths of the CPCA representation are that:

(i) it involves a rather small, very small in comparison to Polynomial Chaos representations, number of parameters, see Table 1.

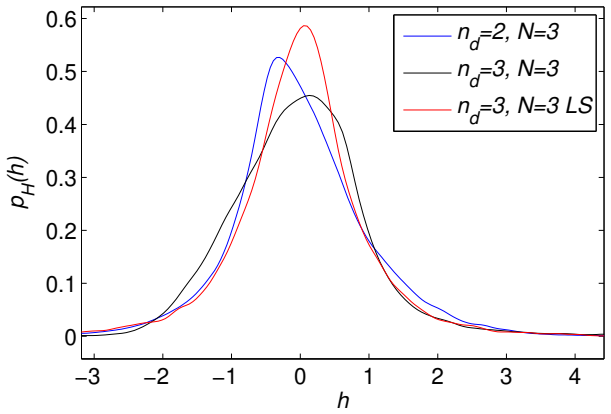

FIG. 11. *Probability density function $p_H(h)$, linear scale, m=15, various optima with direct minimization and least squares approach ("LS").*

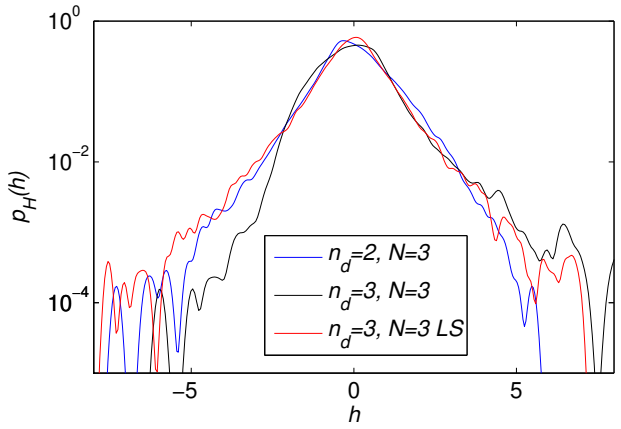

FIG. 12. *Probability density function $p_H(h)$, logarithmic scale, m=15, various optima with direct minimization and least squares approach ("LS").*

(ii) the matching of the distributions of the components X_α as determined from the available data and as modeled from the CPCA representation was found to be good to excellent, depending on the component considered, for the strongly non-Gaussian example considered. Rather interestingly, it was observed that the matching of these two distributions in their tail was consistently very good, an important feature. Such good approximations were obtained with 12 or less parameters z_k.

(iii) the accuracy of the approximation can be simply quantified by the overlap error of Eqs (2.7) and (2.8).

Additional numerical results not presented here for brevity have further shown that the CPCA representation of a Gaussian random vector $\mathbf{X}$ can also be achieved thus suggesting that the modeling approach is widely applicable. Moreover, it was found that identification strategies based on the PCA variables Γ_α, minimizing either an overlap error on those variables or a least squares metric similar to Eqs (4.14)-(4.16) did not lead to a good CPCA approximation. However, the least squares

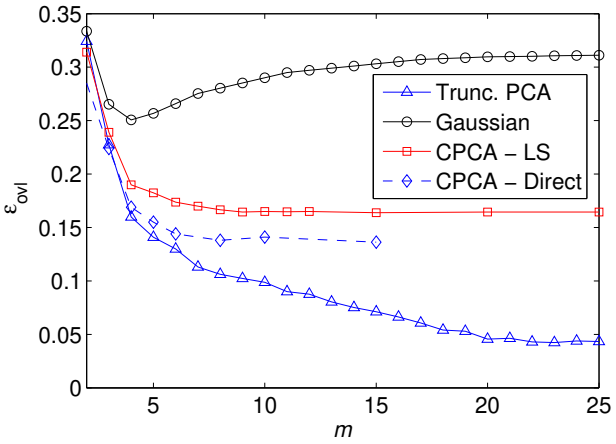

FIG. 13. *Minimum error $\epsilon_{ovl}(N, n_d)$ vs. m for PCA truncated to m=15 terms ("Trunc. PCA"), its Gaussian approximation, and the CPCA optima obtained by direct optimization and least squares approach over the value of N for n_d=3.*

identification approach of Eqs (4.12)-(4.13) was successful, leading to results similar to those presented here in connection with the approach of Eqs (4.14)-(4.16). It is however slightly more expensive than this approach.

Acknowledgments. The financial support of the first author by the Air Force Multi University Research Initiative contract FA9550-15-1-0038 with Drs Jean-Luc Cambier and Fariba Fahroo as Technical Monitors is gratefully acknowledged.

REFERENCES

[1] S. ABRAHAM, M. RAISEE, G. GHORBANIASL, F. CONTINO, AND C. LACOR, *A robust and efficient stepwise regression method for building sparse polynomial chaos expansions*, Journal of Computational Physics, 332 (2017), pp. 461–474, https://doi.org/10.1016/j.jcp.2016.12.015.

[2] M. ARNST, R. GHANEM, AND C. SOIZE, *Identification of bayesian posteriors for coefficients of chaos expansions*, Journal of Computational Physics, 229 (2010), pp. 3134–3154, https://doi.org/10.1016/j.jcp.2009.12.033.

[3] M. BERVEILLER, B. SUDRET, AND M. LEMAIRE, *Stochastic finite element: a non intrusive approach by regression*, European Journal of Computational Mechanics/Revue Européenne de Mécanique Numérique, 15 (2006), pp. 81–92, https://doi.org/10.3166/remn.15.81-92.

[4] G. BLATMAN AND B. SUDRET, *Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach*, Comptes Rendus Mécanique, 336 (2008), pp. 518–523, https://doi.org/10.1016/j.crme.2008.02.013.

[5] G. BLATMAN AND B. SUDRET, *Adaptive sparse polynomial chaos expansion based on least angle regression*, Journal of Computational Physics, 230 (2011), pp. 2345–2367, https://doi.org/10.1016/j.jcp.2010.12.021.

[6] A. BOWMAN AND A. AZZALINI, *Applied Smoothing Techniques for Data Analysis*, Oxford University Press, Oxford, UK, 1997.

[7] B. P. CARLIN AND T. A. LOUIS, *Bayesian Methods for Data Analysis*, Chapman and Hall/CRC, 2008.

[8] B. CHEN-CHARPENTIER AND D. STANESCU, *Parameter estimation using polynomial chaos and maximum likelihood*, International Journal of Computer Mathematics, 91 (2014), pp. 336–346, https://doi.org/10.1080/00207160.2013.809069.

[9] P. CONGDON, *Bayesian Statistical Modelling*, vol. 704, John Wiley & Sons, 2007.

[10] S. DAS, R. GHANEM, AND S. FINETTE, *Polynomial chaos representation of spatio-temporal random fields from experimental measurements*, Journal of Computational Physics, 228

(2009), pp. 8726–8751, https://doi.org/10.1016/j.jcp.2009.08.025.
[11] S. Das, R. Ghanem, and J. C. Spall, *Asymptotic sampling distribution for polynomial chaos representation from data: a maximum entropy and fisher information approach*, SIAM Journal on Scientific Computing, 30 (2008), pp. 2207–2234, https://doi.org/10.1137/060652105.
[12] B. J. Debusschere, H. N. Najm, P. P. Pébay, O. M. Knio, R. G. Ghanem, and O. P. Le Maître, *Numerical challenges in the use of polynomial chaos representations for stochastic processes*, SIAM journal on scientific computing, 26 (2004), pp. 698–719, https://doi.org/10.1137/S1064827503427741.
[13] C. Desceliers, R. Ghanem, and C. Soize, *Maximum likelihood estimation of stochastic chaos representations from experimental data*, International Journal for Numerical Methods in Engineering, 66 (2006), pp. 978–1001, https://doi.org/10.1002/nme.1576.
[14] C. Desceliers, C. Soize, and R. Ghanem, *Identification of chaos representations of elastic properties of random media using experimental vibration tests*, Computational mechanics, 39 (2007), pp. 831–838, https://doi.org/10.1007/s00466-006-0072-7.
[15] A. Doostan, R. G. Ghanem, and J. Red-Horse, *Stochastic model reduction for chaos representations*, Computer Methods in Applied Mechanics and Engineering, 196 (2007), pp. 3951–3966, https://doi.org/10.1016/j.cma.2006.10.047.
[16] O. G. Ernst, A. Mugler, H.-J. Starkloff, and E. Ullmann, *On the convergence of generalized polynomial chaos expansions*, ESAIM: Mathematical Modelling and Numerical Analysis, 46 (2012), pp. 317–339, https://doi.org/10.1051/m2an/2011045.
[17] R. Ghanem, D. Higdon, and H. Owhadi, *Handbook of Uncertainty Quantification*, vol. 1 to 3, Springer, Cham, Switzerland, 2017, https://doi.org/10.1007/978-3-319-12385-1.
[18] R. G. Ghanem and A. Doostan, *On the construction and analysis of stochastic models: characterization and propagation of the errors associated with limited data*, Journal of Computational Physics, 217 (2006), pp. 63–81, https://doi.org/10.1016/j.jcp.2006.01.037.
[19] R. G. Ghanem, A. Doostan, and J. Red-Horse, *A probabilistic construction of model validation*, Computer Methods in Applied Mechanics and Engineering, 197 (2008), pp. 2585–2595, https://doi.org/10.1016/j.cma.2007.08.029.
[20] R. G. Ghanem and P. D. Spanos, *Stochastic Finite Elements: a Spectral Approach*, Springer-Verlag, New York, 1991.
[21] D. Ghosh and R. Ghanem, *Stochastic convergence acceleration through basis enrichment of polynomial chaos expansions*, International journal for numerical methods in engineering, 73 (2008), pp. 162–184, https://doi.org/10.1002/nme.2066.
[22] G. Givens and J. Hoeting, *Computational Statistics*, John Wiley and Sons, Hoboken, New Jersey, 2nd edition ed., 2013.
[23] J. Guilleminot, C. Soize, D. Kondo, and C. Binetruy, *Theoretical framework and experimental procedure for modelling mesoscopic volume fraction stochastic fluctuations in fiber reinforced composites*, International Journal of Solids and Structures, 45 (2008), pp. 5567–5583, https://doi.org/10.1016/j.ijsolstr.2008.06.002.
[24] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, vol. 160, Springer Science & Business Media, 2005.
[25] O. Le Maître and O. M. Knio, *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*, Springer Science & Business Media, 2010.
[26] O. Le Maître, O. M. Knio, H. N. Najm, and R. Ghanem, *Uncertainty propagation using wiener–haar expansions*, Journal of computational Physics, 197 (2004), pp. 28–57, https://doi.org/10.1016/j.jcp.2003.11.033.
[27] D. Lucor, C.-H. Su, and G. E. Karniadakis, *Generalized polynomial chaos and random oscillators*, International Journal for Numerical Methods in Engineering, 60 (2004), pp. 571–596, https://doi.org/doi.org/10.1002/nme.976.
[28] I. MacDonald, *Symmetric functions and Hall polynomials 2nd Ed.*, Oxford University Press, Oxford, UK, 2015.
[29] R. Madankan, P. Singla, T. Singh, and P. D. Scott, *Polynomial-chaos-based bayesian approach for state and parameter estimations*, Journal of Guidance, Control, and Dynamics, 36 (2013), pp. 1058–1074, https://doi.org/10.2514/1.58377.
[30] C. V. Mai and B. Sudret, *Surrogate models for oscillatory systems using sparse polynomial chaos expansions and stochastic time warping*, SIAM/ASA Journal on Uncertainty Quantification, 5 (2017), pp. 540–571.
[31] Y. M. Marzouk and H. N. Najm, *Dimensionality reduction and polynomial chaos acceleration of bayesian inference in inverse problems*, Journal of Computational Physics, 228 (2009), pp. 1862–1902, https://doi.org/10.1016/j.jcp.2008.11.024.
[32] Y. M. Marzouk, H. N. Najm, and L. A. Rahn, *Stochastic spectral methods for efficient*

bayesian solution of inverse problems, Journal of Computational Physics, 224 (2007), pp. 560–586, https://doi.org/10.1016/j.jcp.2006.10.010.

[33] H. N. Najm, *Uncertainty quantification and polynomial chaos techniques in computational fluid dynamics*, Annual review of fluid mechanics, 41 (2009), pp. 35–52, https://doi.org/10.1146/annurev.fluid.010908.165248.

[34] A. Nouy and C. Soize, *Random field representations for stochastic elliptic boundary value problems and statistical inverse problems*, European Journal of Applied Mathematics, 25 (2014), pp. 339–373, https://doi.org/10.1017/S0956792514000072.

[35] G. Perrin, C. Soize, D. Duhamel, and C. Funfschilling, *Identification of polynomial chaos representations in high dimension from a set of realizations*, SIAM Journal on Scientific Computing, 34 (2012), pp. A2917–A2945, https://doi.org/10.1137/11084950X.

[36] G. Perrin, C. Soize, D. Duhamel, and C. Funfschilling, *Karhunen–loève expansion revisited for vector-valued random fields: Scaling, errors and optimal basis.*, Journal of Computational Physics, 242 (2013), pp. 607–622, https://doi.org/10.1016/j.jcp.2013.02.036.

[37] B. Puig, F. Poirion, and C. Soize, *Non-gaussian simulation using hermite polynomial expansion: convergences and algorithms*, Probabilistic Engineering Mechanics, 17 (2002), pp. 253–264, https://doi.org/10.1016/S0266-8920(02)00010-3.

[38] B. V. Rosić, A. Litvinenko, O. Pajonk, and H. G. Matthies, *Sampling-free linear bayesian update of polynomial chaos representations*, Journal of Computational Physics, 231 (2012), pp. 5761–5787, https://doi.org/10.1016/j.jcp.2012.04.044.

[39] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo Method*, Second Edition, John Wiley & Sons, New York, 2008.

[40] R. J. Serfling, *Approximation theorems of mathematical statistics*, vol. 162, John Wiley & Sons, 1980.

[41] Q. Shao, A. Younes, M. Fahs, and T. A. Mara, *Bayesian sparse polynomial chaos expansion for global sensitivity analysis*, Computer Methods in Applied Mechanics and Engineering, 318 (2017), pp. 474–496, https://doi.org/10.1016/j.cma.2017.01.033.

[42] C. Soize, *Identification of high-dimension polynomial chaos expansions with random coefficients for non-gaussian tensor-valued random fields using partial and limited experimental data*, Computer methods in applied mechanics and engineering, 199 (2010), pp. 2150–2164, https://doi.org/10.1016/j.cma.2010.03.013.

[43] C. Soize, *A computational inverse method for identification of non-gaussian random fields using the bayesian approach in very high dimension*, Computer Methods in Applied Mechanics and Engineering, 200 (2011), pp. 3083–3099, https://doi.org/10.1016/j.cma.2011.07.005.

[44] C. Soize, *Polynomial chaos expansion of a multimodal random vector*, SIAM/ASA Journal on Uncertainty Quantification, 3 (2015), pp. 34–60, https://doi.org/10.1137/140968495.

[45] C. Soize, *Uncertainty Quantification. An Accelerated Course with Advanced Applications in Computational Engineering*, Springer, New York, 2017, https://doi.org/10.1007/978-3-319-54339-0.

[46] C. Soize and C. Desceliers, *Computational aspects for constructing realizations of polynomial chaos in high dimension*, SIAM Journal on Scientific Computing, 32 (2010), pp. 2820–2831, https://doi.org/10.1137/100787830.

[47] C. Soize and R. Ghanem, *Physical systems with random uncertainties: chaos representations with arbitrary probability measure*, SIAM Journal on Scientific Computing, 26 (2004), pp. 395–410, https://doi.org/10.1137/S1064827503424505.

[48] C. Soize and R. G. Ghanem, *Reduced chaos decomposition with random coefficients of vector-valued random variables and random fields*, Computer Methods in Applied Mechanics and Engineering, 198 (2009), pp. 1926–1934, https://doi.org/10.1016/j.cma.2008.12.035.

[49] J. C. Spall, *Introduction to Stochastic Searh and Optimization.*

[50] I. Sraj, O. P. Le Maître, O. M. Knio, and I. Hoteit, *Coordinate transformation and polynomial chaos for the bayesian inference of a gaussian process with parametrized prior covariance function*, Computer Methods in Applied Mechanics and Engineering, 298 (2016), pp. 205–228, https://doi.org/10.1016/j.cma.2015.10.002.

[51] A. M. Stuart, *Inverse problems: a bayesian perspective*, Acta numerica, 19 (2010), pp. 451–559, https://doi.org/10.1017/S0962492910000061.

[52] A. Tarantola, *Inverse Problem Theory And Methods For Model Parameter Estimation*, vol. 89, SIAM, Philadelphia, 2005.

[53] G. Terrell and D. Scott, *Variable kernel density estimation*, The Annals of Statistics, 20 (1992), p. 1236–1265, https://doi.org/10.1214/aos/1176348768.

[54] C. Thimmisetty, P. Tsilifis, and R. Ghanem, *Homogeneous chaos basis adaptation for design optimization under uncertainty: Application to the oil well placement problem*, Artificial

Intelligence for Engineering Design, Analysis and Manufacturing, 31 (2017), pp. 265–276, https://doi.org/10.1017/S0890060417000166.

[55] R. Tipireddy and R. Ghanem, *Basis adaptation in homogeneous chaos spaces*, Journal of Computational Physics, 259 (2014), pp. 304–317, https://doi.org/10.1016/j.jcp.2013.12.009.

[56] P. Tsilifis and R. Ghanem, *Bayesian adaptation of chaos representations using variational inference and sampling on geodesics*, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 474 (2018), p. 20180285, https://doi.org/10.1098/rspa.2018.0285.

[57] P. Tsilifis and R. G. Ghanem, *Reduced wiener chaos representation of random fields via basis adaptation and projection*, Journal of Computational Physics, 341 (2017), pp. 102–120, https://doi.org/10.1016/j.jcp.2017.04.009.

[58] X. Wan and G. E. Karniadakis, *Multi-element generalized polynomial chaos for arbitrary probability measures*, SIAM Journal on Scientific Computing, 28 (2006), pp. 901–928, https://doi.org/10.1137/050627630.

[59] D. Xiu and G. E. Karniadakis, *The wiener–askey polynomial chaos for stochastic differential equations*, SIAM journal on scientific computing, 24 (2002), pp. 619–644, https://doi.org/10.1137/S1064827501387826.